MULTILEVEL VALIDITY: ASSESSING THE VALIDITY OF SCHOOL-LEVEL INFERENCES FROM STUDENT ACHIEVEMENT TEST DATA

Sharyn L. Rosenberg

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Education

Chapel Hill
2009

Approved by
Advisor: Dr. Gregory J. Cizek
Reader: Dr. Daniel J. Bauer
Reader: Dr. Jeffrey A. Greene
Reader: Dr. Daniel J. Heck
Reader: Dr. William B. Ware

ABSTRACT
Sharyn L. Rosenberg
Multilevel validity:
Assessing the validity of school-level inferences from student achievement test data
(Under the direction of Dr. Gregory J. Cizek)

Psychometric theory is clear about the central role of validity and the importance of gathering evidence for a particular purpose. State achievement tests are generally developed with ample validity evidence for their intended inferences about student achievement. Such evidence may not be sufficient for drawing group-level inferences, a crucial point that is often ignored when student achievement scores are used in multilevel analyses to study effects at the school level. This study explores the process of gathering multilevel validity evidence necessary to make school-level inferences from student achievement tests.

Using data from approximately 28,000 students in grades 3, 5, and 8 in a northeastern U.S. state, this study examined the multilevel factor structure of mathematics achievement tests. Multilevel exploratory factor analyses were used to determine the optimal number of factors at both the student and the school levels of analysis. Multilevel confirmatory factor analyses were used to assess the extent to which the one-factor solutions on each level were feasible. Both standard (single level) confirmatory factor analyses and multilevel confirmatory factor analyses were used to compare the size and relative importance of factor loadings at the different levels of analysis. The statistical significance of the school-level factor loadings provided evidence about the extent to which the mathematics achievement test items were effective for discriminating between schools.

For each of the three grades studied, there was only one meaningful factor identified (presumably mathematics achievement) at both the student and school levels of analysis. At each grade level, items differed in terms of both their absolute and relative size of their factor loadings at the student and school levels of analysis, suggesting that when school-level inferences are of interest, standard factor analyses provide insufficient information about test development and validation. The majority of items in this study were more discriminating at the school level than at the student level. Interpretations of these findings are discussed in the context of relevant research on validity and student achievement. Implications for educational measurement and ideas for future research are also addressed.

ACKNOWLEDGEMENTS

I would like to thank my husband, Dan Rosenberg, for his support throughout this process. His love and encouragement have helped keep me motivated during the long journey from my first graduate class to completing the dissertation. I also appreciate the sacrifices made by Livia and Cassidy, who sometimes could not spend as much time with me as they would have liked while I completed the dissertation.

I owe so much to Dr. Gregory Cizek, the most incredibly supportive advisor not only throughout the dissertation process, but during my entire graduate school career. His steadfast encouragement, support, and respect have done so much to help me maneuver through this program, and I do not know what I would have done without him. I am also thankful for the guidance and technical expertise provided by my readers, Drs. Daniel Bauer, Jeffrey Greene, Daniel Heck, and William Ware. I am indebted to Dr. Marianne Perie for her generous offer to help secure data for this project. I am also grateful for the insights and challenging questions provided by my colleagues at Horizon Research; in particular, to Dr. Sean Smith, whose questions about group-level measurement properties inspired this study.

Some of my most helpful advisors throughout this process were my fellow graduate students, including Samantha Burg, Bonnie Dahlke, Heather Koons, Dorene MacKinnon, and Patricia Sylvester, among many others. I am also very thankful for the love and support provided by my parents, extended family, and friends throughout my graduate school career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

During the past decade, measurable student achievement outcomes have been brought to the forefront of both educational research and policy. State efforts to track student achievement as an indicator of school quality preceded one of the most influential federal education policies of all time, the No Child Left Behind Act of 2001 (NCLB, 2002). In addition to the unprecedented attention and stakes given to school accountability programs, NCLB (2002) has had a vast impact on the types of outcomes that are measured to assess educational interventions and other school-level processes. In the current climate, it would be difficult to obtain funding for research on school processes or for evaluations of school-level programs without intending to connect those processes or programs to student achievement, whether as a primary or secondary purpose.

Contemporary debates about school effectiveness have tended to focus primarily on student achievement, not because behaviors, attitudes, and other student outcomes are deemed unimportant, but for the simple reason that test scores are so prominent now (Rumberger & Palardy, 2004). Similarly, the goal of increasing student achievement has become a primary focus of many educational programs, even those that do not directly serve students. For example, many studies of school-level programs (such as professional development workshops for principals or teachers) have goals of increasing student achievement for participating schools, even though students do not receive the treatment

directly. Consequently, the evaluations of such programs also tend to incorporate student achievement tests as outcome measures.

In addition to the effect that NCLB has had on the prominence of student achievement, the annual administration and standardization of state achievement tests have resulted in a rich cache of available data for secondary purposes in educational research and program evaluation. State achievement tests undergo rigorous psychometric analyses during test development, are administered on a consistent basis, and are uniform across schools within a given state.

The increased focus on student achievement and the availability of state test data have occurred during a time of widespread development and use of software for multilevel analysis in social science research. Multilevel models are used when the data structure is hierarchical, such as when individuals are clustered within groups (e.g., students within schools) or observations are clustered within persons (i.e., repeated measures design). Multilevel analyses can incorporate different variables at each level of the data hierarchy; an outcome can be modeled as a function of both individual and group predictors. Multilevel analyses can be used not only to account for the complex design of students nested within schools, but also to study the effects of a program that is administered at the school level. When the primary effect of interest is a school-level variable (e.g., principal or teacher participation in a program), the between-school variation in student achievement test scores can be used as a measure of school achievement.

The increased emphasis on student outcomes, wealth of available data, and advancements in analytic techniques have all vastly increased the use of student test scores for secondary purposes in educational research and program evaluation. The use of state

achievement tests for these secondary purposes appears to offer many benefits. State testing programs invest considerable resources into the development and administration of student achievement tests, unlike many lesser-quality measures that may lack adequate psychometric properties. However, it is important to recognize that the stated purpose of most state testing programs is to measure the achievement of students rather than schools. Thus, it is imperative to investigate the validity of state achievement tests for the particular purpose of assessing school-level achievement.

CHAPTER 2

LITERATURE REVIEW

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999; hereafter, *Standards*), the most important characteristic of the measurement process is validity. The following sections trace the importance of gathering validity evidence for the specific purpose of using student achievement results to make inferences about school-level outcomes. A discussion of validity in general is followed by an overview of the most common sources of validity evidence. There is also a focus on considerations specific to nested data (students in schools), most prominently, multilevel analyses. The dual goals of gathering appropriate validity evidence and accounting for multilevel data structures, typically independent procedures in current educational research, are woven together to frame the research questions in the context of multilevel validity.

Validity

Validity has been defined by Messick (1989) as, "an integrative, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13). If the inferences made from test scores are not supported by adequate validity evidence, the intended use of the scores is threatened regardless of how high the reliability of the scores.

Validity is a characteristic of an inference made from a test score rather than a characteristic of the test itself. Validity is a matter of degree; inferences are not valid or invalid, but rather, they are supported by validity evidence that ranges from strong to weak, confirming to disconfirming (Crocker & Algina, 1986). A test can have strong validity evidence for one purpose but little, no, or even contrary evidence for a different purpose. In the chapter on validity, the *Standards* (AERA, APA, & NCME, 1999) emphasizes the importance of gathering validity evidence for each intended use of a test. Such a practice is fundamental to responsible measurement and is essential for ensuring test fairness and preventing test misuse.

The final chapter of the *Standards* (AERA et al., 1999) focuses on the use of tests in program evaluation and public policy. A primary consideration in using tests for program evaluation is the focus on measuring groups rather than individuals (AERA et al., 1999). Standard 15.1 highlights the importance of gathering validity evidence for each intended test purpose and includes the following comment:

> In educational testing, for example, it has become common practice to use the same test for multiple purposes (e.g., monitoring achievement of individual students, providing information to assist in instructional planning for individuals or groups of students, evaluating schools or districts). No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may diminish validity for other purposes (p. 167).

The measurement community acknowledges that psychometric considerations may vary according to the level of measurement (individual vs. group), but little guidance is provided on how to translate this acknowledgment into practice. The *Standards* (AERA et al., 1999) does not include any details or examples about how test validation approaches could be adapted from the measurement of individuals to groups, despite the insistence that it is

important to do so. Nor is there much information on the types of validity evidence most likely to vary according to the level of measurement.

In a recent special issue of *Educational Measurement: Issues and Practice* dedicated to assessing the adequacy of the current *Standards*, Linn (2006) discussed the current lack of guidance on psychometric issues related to aggregate test results. Although he argued that the shortcoming does not necessitate a revised edition of the *Standards,* Linn (2006) proposed several alternatives, including the development of, "a companion set of standards that would specifically address the uses of aggregate test results for evaluation or accountability purposes" (p. 56). Linn (2006) also noted that the lack of guidance on group-level measurement issues is not a new dilemma, but is a challenge that has been present for several decades. For example, the third edition of the *Standards* (AERA, APA, & NCME, 1974) contained a section on using tests for measuring groups in program evaluation, but the revision committee argued that group-level measurement issues were beyond the scope of the *Standards* and would necessitate the development of several different sets of standards (Linn, 2006).

The discussions surrounding the revision of the 1974 *Standards* led to the development of a separate set of standards for program evaluation called *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee on Standards for Educational Evaluation, 1981); the revised (current) edition was published in 1994. The original version of the program evaluation standards included only eight pages on measurement, and a brief introduction to reliability and validity (Joint Committee on Standards for Educational Evaluation, 1981). The revised edition was also not designed to be a comprehensive reference in psychometrics; in fact, Standard A5.C encourages evaluators to

consult the *Standards for Educational and Psychological Testing* for direction on using tests in an evaluation (Joint Committee on Standards for Educational Evaluation, 1994). If program evaluation is considered to fall outside of the traditional boundaries of educational and psychological testing, it is not clear where evaluators should turn for technical expertise on measurement issues that are specific to program evaluation, such as drawing group-level inferences from individual measures.

The most recent edition of *Educational Measurement* (Brennan, 2006) includes a new chapter devoted to group-level measurement issues. However, in that chapter group-score assessments are essentially defined as tests relying on matrix sampling designs, such as NAEP, where group-level inferences are the only (or primary) purpose of the test (Mazzeo, Lazer, & Zieky, 2006). Mazzeo et al. (2006) did not refer to group-level measurement issues in program evaluation, nor to any instances where group-level inferences are a secondary use of tests that were originally developed to make inferences about individuals. In addition, there is no discussion of issues related to test development and validation that are specific to group-level measurement. The only such reference is a justification for the absence of this information, provided in a note that states: "Group-score assessments use item analysis for quality control purposes, and conduct DIF analyses. However, these are not different from approaches used in individual-score tests, so are not discussed here" (p. 694). The chapter on test validation (Kane, 2006) is similarly devoid of considerations related to the measurement of groups.

In a seminal article in psychometrics, Ebel (1961) said of validity: "It is universally praised, but the good works done in its name are remarkably few" (p. 640). Several decades later, Brennan (1998) also referred to the wide gap between validity theory and practice.

Although validity theory in general is hardly impoverished, guidance related to the validation of group-level inferences is lacking. Validity theory related to the measurement of individuals versus groups exists only on the broadest level possible in the form of general emphasis on establishing validity evidence for each intended use of a test. Clear guidance on how test development and validation efforts might differ based on whether the intended score inference pertains to individuals or groups, and advances in this regard are necessary to link validity theory and assessment practice. Without an understanding of how validity considerations may differ according to the level of measurement, it would be surprising if the good works done in the validation of group-level inferences were *not* remarkably few.

Sources of Validity Evidence

According to the *Standards,* current potential sources of validity evidence include: evidence based on test content; evidence based on response process; evidence based on internal structure; evidence based on relations to other variables; and evidence based on test consequences (AERA et al., 1999). Validity evidence based on test content is gathered by consulting with subject matter experts to determine the extent to which the test items adequately sample from the specified domain of interest (AERA et al., 1999; Crocker & Algina, 1986). Content validity evidence is necessary to draw inferences on performance from the sample of test items to the entire domain (Messick, 1989).

In state achievement tests, content validity evidence generally is gathered by comparing test items to state guidelines for the standard course of study in that subject area. The test domain typically is constructed through curriculum analysis, examination of state content standards, and consultation with subject matter experts. Test specifications are used to guide the comprehensive item writing process, where subject matter experts generate an

initial pool of items to represent the target domain. Through the process of field testing and committee review, item content validity is evaluated for accuracy, relevance, and representativeness (Messick, 1989). Alignment studies also may be used to assess the extent to which operational test items are an adequate sample of the content standards in the test domain (Webb, 1999).

Response process analyses seek to gather evidence that test-takers are actually engaging in the processes relevant to the construct being measured. Such evidence is often gathered by questioning examinees about the cognitive processes they engaged in while taking a test, a procedure sometimes referred to as a think-aloud protocol or cognitive interview (AERA et al., 1999; Willis, 2005). In multiple choice tests, cognitive interviews also may provide evidence that examinees are choosing the correct answers for the intended reasons. Depending on the type of test, response process analyses may also include rater judgment, monitoring of records, and analysis of examinee eye-movements or response times (AERA et al., 1999).

Internal structure analyses refer to procedures based on how individual test items relate to each other and therefore how they conform to the intended construct(s). If a test consists of subtests that purport to measure distinct constructs, then internal structural analyses should support the proposed test structure (AERA et al., 1999; Crocker & Algina, 1986). The most common methods for gathering evidence based on internal structure include coefficients of internal consistency (e.g., coefficient alpha), nonparametric approaches to dimensionality, and factor analysis.

Validity evidence by relation to other variables includes several types of evidence. Convergent validity evidence is gathered when a test correlates highly with other measures of

the same or similar constructs. Discriminant validity evidence calls for lower correlations with measures of different constructs. Efforts to gather validity evidence for a mathematics achievement test may include demonstrating higher correlations with other measures of mathematics than with reading comprehension tests. Validity evidence may also be gathered by studying the relationship between a test and a criterion, either present (concurrent) or future (predictive), such as correlations between achievement test results and course grades (AERA et al., 1999; Crocker & Algina, 1986).

The newest and most controversial source of validity evidence is evidence based on test consequences. Proponents of consequential validity evidence argue that both intended and unintended consequences of test use can affect the validity of the test inferences. According to the *Standards,* "evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components" (AERA et al., 1999, p. 16). However, other consequences that are not a direct result of construct underrepresentation would not bear on the validity of the intended inferences.

It seems that construct underrepresentation would pose a threat to validity in general, beyond its impact on consequential validity. Although test consequences from sources other than construct underrepresentation are certainly important to the measurement process, it is not clear that they are appropriately subsumed under validity evidence. Despite calls for consequential validity evidence over the past 20 years, academics and testing professionals do not appear to have embraced this aspect of validity theory. In a review of tests appearing in the current edition of the *Mental Measurements Yearbook* (Spies & Plake, 2005), Cizek,

Rosenberg, and Koons (2008) found that less than three percent of reviews included any reference to consequential validity evidence.

Factor Analysis as Validity Evidence

A common source of validity evidence for internal structure analysis is factor analysis, a statistical technique in which a correlation matrix is used to investigate how measured variables (items or tests) relate to each other. Factor analysis is based on the assumption that high correlations among variables (e.g., test items) are due to a smaller set of common causes, or latent factors. Factor analytic theory was first developed a little over a century ago in relation to intelligence testing, when Spearman (1904) hypothesized that a general ability factor (g) accounted for correlations in performance across multiple tests. The mathematical model was later refined by Thurstone (1947), who developed many of the principles of modern factor analysis, such as rotation and simple structure.

The original applications of factor analysis were examples of exploratory factor analysis, a technique with the purpose of identifying the number and nature of the common factors underlying a set of variables. Confirmatory factor analysis, a technique used to test *a priori* hypotheses about how measured variables relate to underlying factors, was subsequently developed and programmed by Jöreskog (1969). Both exploratory and confirmatory factor analysis are now widely used in educational and psychological research for many different purposes.

Although factor analysis was not originally developed for the purpose of gathering validity evidence, the technique has often been used for this purpose. Factor analysis is a method of gathering validity evidence because test items that were created to measure the same construct should be moderately correlated with each other and relate to the same

underlying factor (Crocker & Algina, 1986). Although exploratory factor analysis can be very helpful during test development, confirmatory factor analysis provides stronger test validation evidence because it directly addresses the question of whether the test appears to be measuring what it was intended to measure.

In a review of the role of factor analysis in test validation, Goodwin (1999) traced one of the first references of the practice to the 1966 edition of the *Standards.* Interestingly, this early mention of using factor analysis as validity evidence was in the form of a caution that such evidence alone is insufficient and recommended that concurrent validity evidence was needed: "A new interest test that emphasizes the factorial approach to construct validity should nevertheless report relationships of the new instrument to relevant scales of some well-established tests" (APA, 1966, p. 23-24; as cited in Goodwin, 1999, p. 92).

The current version of the *Standards* is much less judgmental in terms of the adequacy and importance of specific validity evidence requirements, although the *Standards* do emphasize the need to integrate several different sources of evidence (AERA et al., 1999). Reference to factor analysis in the *Standards* is made indirectly in the form of evidence related to the internal structure of tests: "Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 1999, p. 13). In the most recent edition of *Educational Measurement*, Kane (2006) cautioned that patterns in test performance may be due to something other than the hypothesized constructs. Although factor analysis alone may not yield sufficient evidence to support a specific test use, a lack of factorial validity may seriously threaten a proposed inference. A

lack of factorial validity would pose a serious threat to the intended inferences, even if other sources of validity evidence (such as evidence based on test content) supported the test use.

Messick (1989) asserted that test developers should avoid the temptation to focus validation efforts on claims that are easy to support; rather, they have the responsibility to investigate rival hypotheses that legitimately threaten the proposed test score interpretations. The unitary concept of validity (Messick, 1989) and the argument-based approach (Kane, 1992) are not intended to be used as a means of cherry-picking the types of evidence that are most likely to yield confirming evidence. Camara and Lane (2006) noted that the removal of judgmental language in the last revision of the *Standards*, "appears to have moved the current standards to a more aspirational level" (p. 36). Unlike previous editions that deemed certain types of evidence as "Essential" (APA, 1954) or "Primary" (AERA et al., 1985) and largely placed the burden of gathering this evidence on test developers, the current *Standards* has, to some extent, shifted to test developers and/or users the responsibility to identify and investigate the most important threats to validity. If a test interpretation rests on the assumption that items are related to each other in a particular fashion, then factor analytic evidence seems imperative.

Factor Analysis with Ordered-Categorical Variables

Traditional factor analytic techniques assume that measured variables (items or tests) are continuous and have an interval or ratio scale, an assumption that is clearly violated with dichotomous test data. If the categorical nature of the measured variables is ignored and traditional factor analytic techniques are applied, factor loadings will be biased and their standard errors will be underestimated. Both theory and software have been developed to

adapt linear factor analytic techniques to ordered-categorical variables (including binary data), a technique also known as item factor analysis (McDonald, 1999).

One variation of item factor analysis that is implemented in some current factor analytic software (e.g., LISREL, Jöreskog & Sörbom, 2006; MPLUS, Muthén & Muthén, 2008) is the underlying variable approach, where ordered-categorical variables are posited to represent an unobserved continuous variable. For example, a mathematics test item is either correct or incorrect, but it may represent the continuous latent variable of mathematics ability. Factor analytic models with ordered-categorical variables also produce thresholds that divide the categories. The number of thresholds produced is one less than the number of categories, with binary items having only one threshold that represents the location on the underlying variable that corresponds with a 50% probability of endorsing the item (in the context of surveys or attitude scales) or of answering the item correctly (in the context of dichotomously-scored achievement tests). These models are estimated by computing tetrachoric or polychoric correlations and using a method of weighted least squares estimation.

Cluster Sampling

Most analytic techniques (including factor analysis) are based on the assumption that data were obtained from a simple random sample. In a simple random sample, all population members have an equal probability of inclusion in the sample and all possible samples are equally likely to occur (Lohr, 1999). Random number tables (or computer programs) generally are used to select simple random samples. In education, a simple random sample of students in the United States could be chosen by selecting students from across the nation without regard to their schools; this would likely result in a sample that included very few

students per school, scattered across many different locations. Such an approach is often not practical or cost effective in educational research. In practice, a true simple random sample is rare; cluster sampling is much more common, where individuals occur in related groups such as schools. Consequently, the cluster grouping contributes an additional source of variation to the sample. For example, if a multi-stage sample is drawn where a group of schools is chosen, followed by selection of some or all students in those schools, it is likely that students within the same school have more in common with each other than a true simple random sample of students across all possible schools. The issue surrounding the measurement of individuals nested in higher level groups has long been recognized in education and was once known as the "unit of analysis problem" (Knapp, 1977).

Cluster samples are less statistically efficient than simple random samples; that is, they usually require a much larger sample size to achieve the same level of statistical precision as a simple random sample (Lohr, 1999). The *design effect* is a ratio of the variance from the actual sampling plan to the variance of a simple random sample comprised of the same number of units (Lohr, 1999). The size of the design effect is largely dependent on the intraclass correlation coefficient (ICC), a measure of how similar the members of a cluster are (Lohr, 1999). The ICC and design effect are relatively large when cluster members are fairly homogeneous. In cluster sampling, the design effect is often much larger than one, indicating that many more sample members are required to achieve the same degree of statistical precision attained by a simple random sample. Consequently, when the design of a cluster sample is ignored during analysis and instead is treated as a simple random sample, parameter estimates are biased and standard errors are underestimated (Lohr, 1999).

There are two divergent approaches to addressing the effects of cluster samples, where individuals are nested in higher level groups. If the group-level variation is viewed as a nuisance (that is, as something that must be corrected in order to obtain accurate results for the overall sample), then a complex sampling approach is used. The complex sampling approach would be preferable if the primary purpose of the analysis was to produce estimates that are applicable to all students, regardless of their schools. Complex sampling programs such as SUDAAN (Research Triangle Institute, 1994), WesVar (Westat, 2000), or MPLUS (Muthén & Muthén, 2008) can be used to account for the clustered nature of the data by producing accurate point estimates and standard errors. The second alternative to working with cluster samples is multilevel modeling. This alternative is appropriate when the group-level variation is not considered to be "noise" to be overcome but relates to questions of theoretical interest. The multilevel modeling approach would be preferable if the purpose of the analysis was to consider sources of variation at both the individual and school levels. There are many available computer programs for performing multilevel analyses, including HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), MPLUS (Muthén & Muthén, 2008), SAS (SAS Institute, 2005), SPSS (SPSS Inc., 2006) and MLwiN (Goldstein, 1998).

Multilevel Considerations

The multilevel approach is increasingly common in educational research because sources of variation related to the teacher, classroom, or school are often of theoretical or practical interest. The application of multilevel methods to education datasets is so natural, in fact, that many of the didactic references on multilevel analysis are based on data from students and schools (Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Singer, 1998). The multilevel approach enables researchers to ask new questions about

variables at higher levels of aggregation, such as effects related to schools. Previously, such questions could only be addressed by using a single school-level mean as the outcome variable, but this practice ignores within-school variation and leads to overestimates of statistical precision (Raudenbush & Bryk, 2002).

In educational research, multilevel analysis has become a widely accepted alternative to traditional regression approaches, due to advances in both analytic techniques and software development. There are many examples of multilevel analyses both within and outside of education, particularly within the past few years. In multilevel modeling, an outcome can be modeled as a function of different predictor variables on multiple levels of analysis. For example, consider a two-level model, where students (level one observation units) are nested within schools (level two observation units). A student achievement outcome can be modeled as a function of student-level characteristics such as demographics and prior achievement (level one variables) and school-level characteristics such as school size, school proportion of free/reduced lunch students, and school participation in an intervention (level two variables).

Although it is not as ubiquitous in research as traditional HLM, multilevel structural equation modeling (including multilevel factor analysis) has also been developed to incorporate latent variables into multiple levels of analysis (Heck, 2001; Heck & Thomas, 2000; Muthén, 1991; Muthén, 1994; Muthén & Satorra, 1995). The theory is based on decomposing the total covariance matrix into separate components for between-level and within-level variation (Muthén, 1991). Application of multilevel structural equation modeling has demonstrated that relationships between measured and latent variables are not

necessarily the same across different levels of analysis (Heck, 2001; Heck & Thomas, 2000; Muthén, 1991; Muthén, 1994; Muthén & Satorra, 1995).

There are several examples in educational research where multilevel factor analysis has been applied and a different factor structure was supported at each level. For example, Härnqvist, Gustafsson, Muthén, and Nelson (1994) analyzed verbal and numerical ability data at the individual and class levels for students in grades four through nine and found that several factors that appeared at the individual level were not supported at the class level. Using intelligence data from Van Peet (1992), Hox (2002) performed a multilevel factor analysis for students nested within families and found that although both verbal and numerical factors were supported at the student level, only a general intelligence factor could be extracted at the family level. In a multilevel factor analysis of speaking and writing, Kuhlemeier, van den Bergh, and Rijlaarsdam (2002) found seven speaking factors at the student level (corresponding to different types of speaking situations) but only a single general speaking factor at the school level. These examples support Muthén's (1989) finding that the number of factors at the within level generally serves as an upper limit to the number of factors that can be extracted at the between level.

In addition to the use of multilevel analyses, nested data structure is recognized as an important consideration during the study design phase. Shadish, Cook, and Campbell (2002) discussed the importance of designing experimental studies where the unit of assignment matches the unit of analysis. Researchers are cautioned not to conduct experiments where the effect of interest is confounded with the aggregate unit, such as assigning each teacher or school to a different experimental condition (Shadish, Cook, & Campbell, 2002; What Works Clearinghouse, 2006). Nested data structure is also an important consideration in power

analyses; several software programs have been development to incorporate information about the unit of analysis, including Optimal Design (Raudenbush & Liu, 2000) and PINT (Snijders & Bosker, 1993). The design effect is often used to inform requirements for sample size and power.

Although the psychometric implications of the unit-of-analysis problem (Sirotnik, 1980) have been known for over 25 years, issues in multilevel measurement have received much less attention in research and applied work than considerations in analysis and design. Sirotnik (1980) argued that issues related to the unit of analysis are almost never accounted for during the psychometric phase of research, even when the partitioning of effects into multiple levels is a primary goal of analysis. Sirotnik credited Cronbach (1976) with raising this issue in regard to the aptitude-by-treatment interaction: "Once the question of units [of analysis] is raised, all empirical test construction and item-analysis procedures need to be reconsidered" (Cronbach, 1976, p. 9.19-9.20; as cited in Sirotnik, 1980, p. 249). Psychometric considerations are important because the factor structure at different levels can yield conflicting results; therefore, it is crucial to coordinate test development and validation efforts with the appropriate analyses (Sirotnik, 1980).

Not all multilevel models necessitate using the between level of analysis for test development and validation; this determination depends on the study purpose. If the primary effects and outcomes are conceptualized at the individual level, such as studying effects of gender on student achievement, then the within level is most appropriate. Multilevel factor analysis is still an appropriate technique when students are nested within schools, but the pooled within-level covariance matrix would be used for the effects of interest. If the total covariance matrix was used (as is the case in a traditional factor analysis), effects related to

both levels would be confounded. Unless the school-level variation is high, the total covariance matrix is likely to be more influenced by the pooled within-level covariance matrix, so the conclusions may not be that different (Sirotnik, 1980).

If the main effect of interest is at a higher level of aggregation, however, then the between-level covariance matrix is most appropriate (Hox, 2002). For example, school participation in a program is a level two variable in a model with students and schools, so the student achievement outcome is being used to study systemic effects on school achievement. In this scenario, the between-level variation is most important to theoretical considerations. Use of the total covariance matrix would be incompatible with the study purpose, because it is possible for the between-level variation to be quite different from the total or pooled-within variation. It is very possible for decisions related to test development and validation to differ depending on whether the total, pooled-within level, or between level covariance matrices were used for the analyses.

Multilevel Factor Analysis for Test Development or Validation

Although issues related to multilevel psychometrics were raised over 25 years ago, it is still very rare for test development and validation efforts to consider multiple levels of analysis. There are a few recent examples in the literature where multilevel structural equation modeling has been used for test development or validation by examining psychometric properties at multiple levels of analysis. The examples span many disciplines; Table 2.1 provides a list of several recent examples of the use of multilevel SEM in psychometric analyses by the area of application and corresponding citation.

Table 2.1

Recent Examples of Multilevel Structural Equation Modeling in Psychometric Analyses

Area of application	Citations
Business	Cheung, Leung, and Au (2006); Dyer, Hanges, and Hall (2005); Hall, Hanges, and Dyer (in press); Van de Vijver and Watkins (2006); Zyphur, Kaplan, and Christian (2008)
Education	Allodi (2002); Branum-Martin, Mehta, Carlson, Carlo, Fletcher, Ortiz, and Francis (2006); Farmer (2000); Janus and Offord (2007); Kaplan and Elliott (1997); Kaplan and Kreisman (2000); Kuhlemeier, van den Bergh, and Rijlaarsdam (2002); Mehta, Foorman, Branum-Martin, and Taylor (2005); Toland and De Ayala (2005); Van Horn (2003); Zimprich, Perren, and Hornung (2005)
Health Care	Reise, Meijer, Ainsworth, Morales, and Hays (2006); Sexton, Helmreich, Neilands, Rowan, Vella, Boyden, Roberts, and Thomas (2006); Zhang and Wan (2005)
Neighborhoods	Cerin, Leslie, Owen, and Bauman (2008); Cerin, Saelens, Sallis, and Frank (2006)
Psychology	Li, Duncan, Duncan, Harmer, and Acock (1997); Papaioannou, Marsh, and Theodorakis (2004); Reise, Ventura, Nuechterlein, and Kim (2005)
Sociology	Steele and Goldstein (2006)

In the field of education, multilevel structural equation modeling has been used to conduct psychometric analyses on language and literacy, school readiness, school climate, self-esteem, student evaluations of teaching, and education indicators. Although Muthén (1991) conducted analyses on student and class components of mathematics achievement subtests, no examples with student achievement tests could be found with item-level data or with the purpose of investigating psychometric issues related to school-level achievement.

Research Questions

It is becoming increasingly common to use student achievement tests for the secondary purpose of measuring school achievement, particularly when the main effect of interest occurs at the school level. Professional standards are clear that intended, secondary uses of tests require additional validity evidence, but there is a lack of guidance on how to investigate potential differences in evidence for individual and group-level inferences. Multilevel factor analysis presents an opportunity to examine the internal structure of student achievement tests at both the student and school levels. The possibility for the factor structure to differ at multiple levels of analysis presents a plausible threat to the validity of school-level inferences. It is incumbent upon test users to investigate such threats before using student achievement data to make inferences about school achievement.

This study seeks to address the following general question: Can analysis of the multilevel factor structure of large scale educational achievement tests provide validity evidence for drawing school-level inferences? State mathematics achievement test data from all students in grades 3, 5, and 8 in a single state were used in the study.

The specific research questions are as follows:

1. What is the optimal factor structure at each level of analysis?
2. To what extent is a one-factor solution feasible at both the within (student) and between (school) levels of analysis?
3. How do factor loadings of a one-factor solution differ on the within (student) and between (school) levels of analysis? How do these loadings compare to an overall analysis where data are collapsed across levels?
4. Are the factor loadings at the between level significantly different from zero?

Summary

Psychometric theory is clear about the central role of validity and the importance of gathering evidence for a particular purpose. State achievement tests are generally developed with ample validity evidence for their intended inferences about student achievement. Such evidence may not be sufficient for drawing group-level inferences, a crucial point that is often ignored when student achievement scores are used in multilevel analyses to study effects at the school level. This study explores the process of gathering multilevel validity evidence necessary to make school-level inferences from student achievement tests.

CHAPTER 3

METHODS

The study used secondary data from a statewide student achievement testing program in a northeastern state in the United States to assess the extent to which inferences about school-level achievement are supported. The dataset consisted of all students who took the Spring 2006 mathematics state achievement test in grades 3, 5, and 8. The study participants, performance measures, and data analysis procedures are described below.

Participants

The total sample consisted of all 28,200 students in grades 3, 5, and 8 who participated in the Spring 2006 state achievement tests for mathematics. Descriptive statistics on the number of schools and number of students per school at each grade level are presented in Table 3.1.

Table 3.1

Number of Students per School by Grade

Test	Schools	Number of Students per School			
		Minimum	Maximum	Mean	SD
Mathematics Grade 3	111	1	180	80.95	39.82
Mathematics Grade 5	96	1	395	93.51	77.85
Mathematics Grade 8	84	1	541	121.88	145.74

Student demographic data are presented in Table 3.2. At each grade level, the percentage of males and females was approximately equal and slightly more than half of all students were Caucasian.

Table 3.2

Percent Gender and Ethnicity by Grade

	Gender		Ethnicity[1]				
Test	Male	Female	W	AA	A	H	AI
Mathematics Grade 3 (N=8,985)	51	49	53	33	3	11	0
Mathematics Grade 5 (N=8,977)	52	48	53	34	3	10	0
Mathematics Grade 8 (N=10,238)	51	49	54	35	3	9	0
Total (N=28,200)	51	49	53	34	3	10	0

[1]W-White; AA-African American; A-Asian; H-Hispanic; AI-American Indian

Measures

The student achievement measures used were a sample of statewide, mandated tests designed to measure achievement in English language arts, mathematics, science, and social studies in grades 2-11. Only the mathematics test data were used for this study. The mathematics achievement tests were designed according to specifications developed by the state's department of Education and were intended to align to state mathematics content standards in each grade. The mathematics achievement tests included selected items from the Stanford Achievement Test, 10th edition (SAT10) in addition to the items created by state educators to measure the state content standards. This merger of state-developed items with commercially produced (or "off the shelf") items, known as an augmented achievement test

(AAT), is becoming increasingly common in state testing programs (Cizek, 2008). The mathematics achievement tests were intended to measure the following four mathematics content strands (i.e., sub-areas): numeric reasoning, algebraic reasoning, geometric reasoning, and quantitative reasoning.

The state mathematics achievement items included several item formats: multiple choice, short answer, and extended constructed response. The majority of each test consisted of four-option multiple choice items. Short answer (SA) questions were scored on a scale of 0-2. Extended constructed response (ECR) questions were scored on a scale of 0-4. Each mathematics achievement test also included a small number of embedded field test items and a few additional SAT10 items that did not count towards a student's actual score (and subsequently were not used in these analyses). The specifications of the operational mathematics achievement test items (including the SAT10 items) are listed in Table 3.3.

Table 3.3

Item Formats for Operational Items by Grade

	Number of Items per Format[1]			
Test	MC (0-1)	SA (0-2)	ECR (0-4)	Total
Mathematics Grade 3	45	14	0	59
Mathematics Grade 5	48	8	3	59
Mathematics Grade 8	28	8	3	59

[1]MC-multiple choice; SA-short answer; ER-extended constructed response
Range of possible points indicated in parentheses

New items were developed according to the test specifications and were reviewed by the state's Test Development Committee for accuracy, alignment to state content standards, and generally accepted testing practices. Any new items appearing on the spring 2006 tests were field-tested the previous year by embedding them in the operational tests. As part of the field testing procedures, all items were reviewed by content experts in addition to bias and sensitivity committees. The Mantel-Haenszel (MH) procedure was used to investigate potential differential item functioning (DIF) by race and gender.

Multiple choice items were scored electronically. Short answer and extended constructed response items were scored by trained raters who were college-educated, attended an intensive workshop specific to this administration of the test (including anchor papers and training sets), and were monitored for accuracy and consistency. Each student response was scored by one trained rater, and 10% of responses were checked for accuracy by a team leader.

Internal consistency reliability measured using Cronbach's alpha was 0.95 for grade 3 mathematics, 0.91 for grade 5 mathematics, and 0.93 for grade 8 mathematics. Correlations between the item types (SAT10, multiple choice, short answer, and extended constructed response) were all moderate to high; correlations ranged from 0.77 to 0.82 for grade 3 mathematics, 0.64 to 0.80 for grade 5 mathematics, and 0.69 to 0.87 for grade 8 mathematics.

Data Analysis

The study analyzed the factor structure of state mathematics achievement data at the student level, school level, and collapsed across both levels. The analysis plan included descriptive analyses, multilevel exploratory factor analyses, multilevel confirmatory factor analyses, and standard confirmatory factor analyses (collapsed across both levels). Data

preparation and descriptive analyses were performed in SPSS version 14.0. Factor analyses and multilevel factor analyses (both exploratory and confirmatory) were performed using MPLUS version 5.2 (Muthén & Muthén, 2008). Because the test items were unique to each grade level, all analyses were performed separately by grade. That is, each step of the analysis plan was conducted three times, once for each grade 3, 5, and 8.

The first step in the analysis plan was to determine the optimal factor structure at each level of analysis (Research Question 1). The purpose of Research Question 1 was to provide background on the structure of the current measures. The general purpose of this study was not to propose a series of complex alternative models, but rather to investigate the extent to which a one factor model is feasible with a focus on informing future test development efforts. The current scoring and analyses of the state assessment data rest on the assumption of a single factor regardless of whether this is the optimal solution. Many item response theory models typically used in state achievement testing programs assume unidimensionality at the student level, and performing school-level analyses effectively assumes unidimensionality at both levels of analysis. Prior to investigating (in subsequent research questions) the extent to which each measure is unidimensional, it was paramount to explore whether a more complex factor structure was appropriate when the data were partitioned into the within (student) and between (school) levels of analysis.

This first research question was addressed using multilevel exploratory factor analyses. In MPLUS version 5.2 (Muthén & Muthén, 2008), multilevel exploratory factor analysis can be performed as a single step, eliminating the need for the user to conduct separate analyses of the pooled-within groups, between-groups, and collapsed covariance matrices as was previously required with Muthén's (1994) four steps for multilevel

exploratory factor analysis. The multilevel exploratory factor analyses were conducted using weighted least squares means (WLSM), where parameter estimates are produced using a diagonal weight matrix, and standard errors and mean-adjusted chi-square statistics are produced using a full weight matrix (Muthén & Muthén, 2007). Although maximum likelihood estimation can also be used to estimate these models, Asparouhov and Muthén (2007) found that for two-level factor analyses with categorical variables, WLSM was superior to maximum likelihood in terms of convergence, robustness, and quality of estimation in MPLUS version 5.2 (Muthén & Muthén, 2008).

Numerical integration was used with an EM algorithm, seven integration points per dimension, and a convergence criterion of 0.001. A probit link was used. Geomin, a type of oblique rotation, was employed to allow the factors to correlate.

The multilevel exploratory factor analyses yielded the following information for evaluating model fit: eigenvalues for within-level correlation matrices and between-level correlation matrices; chi-square tests; CFI; TLI; RMSEA; and SRMR for the within and between levels of analysis. In addition, the intraclass correlation (ICC) of each item provided a descriptive measure of the amount of variation at each level of analysis.

The eigenvalues for the within-level correlation matrices and the between-level correlation matrices were used to construct scree plots (Cattell, 1966) for each level of analysis. Scree plots provide a visual indication of how many factors may be feasible in exploratory factor analysis. The point at which a line drawn through the eigenvalues changed slope was used to provide a rough estimate of the number of factors present on the student level and the school level (Tabachnick & Fidell, 2001).

A summary of the fit indices and criteria used for evaluating the factor structure of the models is presented in Table 3.4.

Table 3.4

Fit Statistics Used to Evaluate the Optimal Number of Factors at Each Level of Analysis

Statistic	Cutoff for Adequate Fit
Chi-square (χ^2)	$p \geq .05$
Tucker-Lewis Index (TLI)	$\geq .95$
Comparative Fit Index (CFI)	$\geq .95$
Root Mean Squared Error of Approximation (RMSEA)	$\leq .06$
Standardized Root Mean Squared Residual (SRMR)	$\leq .08$

The chi-square test is a fit index that evaluates a specified model by comparing it to a model that is just-identified (Kline, 1998). A statistically significant result indicates that the specified model does not fit the data as well as an unrestricted model; thus, p-values of greater than .05 are desirable. However, the chi-square test has been criticized as an inadequate measure of fit because it is overly sensitive to sample size (Kline, 1998). With large samples, it is virtually impossible to find an over-identified model that is not statistically significant. Small sample sizes are much more likely to yield non-significant chi-square values, due to a lack of power for detecting model misfit.

The Tucker-Lewis index (TLI) and comparative fit index (CFI) are incremental fit indices. The TLI (Tucker & Lewis, 1973) compares the fit of a specified model to both a null

model (where there are no factors and all dependent variables are unrelated) and an ideal model (where fit is exact in the population). The CFI (Bentler, 1990) is a non-centrality measure that also compares the fit of the proposed model to a null model. Hu and Bentler (1999) suggested that values of .95 or higher indicate good fit for both the TLI and CFI.

The root mean squared error of approximation (RMSEA; Steiger & Lind, 1980) is a test of close fit; it is a measure of discrepancy that accounts for model complexity by including the degrees of freedom in the denominator. Consequently, increasing the number of parameters in the model will only improve the RMSEA if the decrease in the discrepancy function compensates for the loss of degrees of freedom. Hu and Bentler (1999) recommended a value of .06 or lower for the RMSEA.

The standardized root mean squared residual (SRMR; Bentler, 1995) is based on the covariance residuals. The SRMR is the only fit index computed separately at the within and between levels of analysis. Hu and Bentler recommended values of .08 or lower for the SRMR.

It should be noted that research on the criteria for fit indices has been based on standard (single level) structural equation modeling. It is not clear whether some of the guidelines may differ for multilevel structural equation modeling in general, and particularly for multilevel structural equation modeling with categorical variables.

Finally, the factor structure and pattern of loadings (including size and presence of cross-loadings) were examined to determine whether each solution was interpretable. The unrestricted (just-identified) model at each level was also used to better understand the optimal factor structure at the other level (Muthén & Asparouhov, 2009). All of these

indicators were synthesized to arrive at the judgment of how many factors were optimal at each level of analysis.

Multilevel confirmatory factor analyses with one factor on each level were performed to address research questions 2 through 4. The analytic procedures for the multilevel confirmatory analyses were the same as those used in the multilevel exploratory factor analyses. That is, the models were estimated using weighted least squares means and a probit link function. Numerical integration was performed with an EM algorithm, using seven integration points per dimension and a convergence criterion of 0.001. To produce estimates for all factor loadings, the models were identified by setting the variance to one (on each level) rather than the first factor loading. Factor loading estimates were fully standardized. To determine whether judgments of item quality would differ using the unstandardized factor loadings, a subset of the analyses were re-run using an alternative identification method where the first factor loading was set to one on each level of analysis.

The purpose of Research Question 2 was to investigate the adequacy of a one-factor solution, regardless of whether a more complex factor structure might provide a more optimal solution. It is possible that a unidimensional solution on collapsed data would not hold up when the data were separated into the within and between levels of analysis. Factor loadings on both levels were examined to determine whether any of the loadings were close to zero or negative. The presence of such loadings on one or more levels would indicate that some items were not contributing much to the total test score (in the case of loadings close to zero) or were even detracting from the total test score (in the case of negative loadings). The presence of several low or negative loadings would suggest that a one-factor solution was not

very feasible at a particular level of analysis. The fit statistics listed in Table 3.4 also were used to assess the adequacy of the solution with one factor on each level.

To compare factor loadings across different levels of analysis (Research Question 3), standard factor analyses (ignoring the clustered nature of the data) were performed in addition to the multilevel confirmatory factor analyses. The comparisons of the loadings from the single level factor analyses to the within- and between-level loadings were purely descriptive, as there are no significance tests available for this procedure. The following questions were considered as part of the descriptive analyses: Are the same items contributing the most to the total test score on different levels of analysis? Are the items that appear to be most strongly related to the overall test in a standard factor analysis still strong indicators on one or both levels when the data are separated into the within and between levels of analysis? The purpose of this research question was to determine whether judgments of item quality vary according to the level of analysis.

To compare factor loadings across the within and between levels of analyses, scaled chi-square difference tests were used. The scaled chi-square difference tests evaluated whether models where the factor loadings were constrained to be equal across levels fit significantly worse than models where the factor loadings were estimated freely (independently) at the within and between levels of analysis. Chi-square difference tests for normal outcomes can be performed by taking the difference between the chi-square values of nested models and evaluating the significance using the difference in the degrees of freedom (Kline, 1998). This traditional approach is not appropriate for non-normal outcomes where the Satorra-Bentler scaled chi-square statistic (Satorra & Bentler, 1994) is used, however, because the difference between two scaled chi-square values does not follow a scaled chi-

square distribution. In order to perform the chi-square difference tests, it was necessary to use scaling correction factors produced in the output of the multilevel confirmatory factor analyses, based on formulas produced by Satorra (2000). Models were re-estimated using MLR, a type of maximum likelihood, and the chi-square difference tests were performed using loglikelihood values. The results of the scaled chi-square difference tests were evaluated at a significance level of $p < .05$.

The between-level factor loadings were the focus of Research Question 4. This research question expanded the descriptive analyses performed in Research Question 2 by examining the statistical significance of the loadings at the school level. Wald tests (Tabachnick & Fidell, 2001) were performed by dividing each parameter estimate (factor loading) by the standard error of the parameter estimate. Since this procedure follows a z-distribution, a value greater than 1.96 was judged statistically significant at $p < .05$. Both the statistical significance and direction of the factor loadings were noted. The presence of several non-significant or negative loadings at the between level would suggest that the test items may not be effective for discriminating between schools, regardless of their ability to discriminate between students.

Methodological Limitations

The data analysis procedures have some important limitations. First, the two-level structure of students within schools did not consider class-level variation. This is largely a practical limitation, as state achievement databases typically either do not capture or highly restrict access to teacher information. The software used in this study also imposed such a limitation, as MPLUS (Muthén & Muthén, 2008) and most other alternatives for performing multilevel factor analysis allow for only two levels of analysis in the study procedures.

A second technical limitation is the assumption of homogeneity of the within-groups covariance matrix (Muthén, 1994). That is, using multilevel factor analysis to address the research questions assumes that the factor structure of test items at the student level is the same in every school in the state. Although a handful of multilevel validity studies in the area of cross-cultural research (e.g., Cheung, Leung, & Au, 2006) have performed multilevel measurement invariance analyses to evaluate the homogeneity of the within-groups covariance matrix, such a step is only possible when the number of level two units is relatively small, as occurs when culture is the level two unit. In analyses with large numbers of level two units, such as schools, it is generally not feasible to test this assumption. Although there is no reason to expect that this is a major problem, it must be acknowledged that violation of the assumption of homogeneity of the within-groups covariance matrix could potentially distort results of multilevel factor analyses.

Because the data were from a small state, the analyses were based on a relatively low number of level two units (ranging from 84 to 111 schools). Although the data still met the proposed minimum standard for multilevel factor analysis of 50-100 groups (Muthén, 1994), it is likely that more stable solutions could be achieved with larger numbers of schools. The number of level two units is particularly important for estimating the parameters on the school level. The ratio of the number of schools to the number of between level factor loadings was small, which has the potential to affect model identification, standard errors, and parameter estimates. If data from a larger state had been used instead, however, there would have been a greater chance of models not running due to computational complexity. Additional research using monte carlo simulations would provide guidance about the impact of a small number of level two units relative to the number of estimated parameters.

Another potential issue related to sample size is the presence of small clusters (schools with few students at a particular grade level). As indicated in Table 3.1, there were a few schools with only one student at a particular grade. This may have been due to students testing off-grade or being homeschooled, among other possibilities. Muthén (2002) advised retaining all schools in this situation, even those with only one student. The small clusters still contribute to the between level estimation, even though they are not involved in the estimation of within level parameters. Maas and Hox (2005) studied the effects of sample size at each level in multilevel modeling and concluded that small cluster sizes did not lead to biased parameters or standard errors. The Maas and Hox (2005) study considered only equal-sized clusters, however, where the smallest grouping was five people per cluster. Because there was no compelling evidence to eliminate schools with only one student, and considering the small level two sample size, all schools were retained in the analyses. However, there is no known information about the potential effects of including one person clusters in these types of analyses. More research is needed to determine what guidelines, if any, should be used in eliminating small level two units. To examine the potential impact of the small clusters, multilevel confirmatory factor analyses were re-run using only clusters of five or more, and the results were compared to those produced from the full sample.

A more general limitation of this study is the novelty of the methodology. Multilevel factor analysis for categorical variables is in its infancy in the field of quantitative methods, even in relation to traditional procedures for factor analysis, categorical variables, and multilevel models. Computational knowledge and guidance related to multilevel factor analysis for categorical variables is very limited, and much of the available expertise exists in nontraditional sources such as online message boards (e.g., www.statmodel.com) rather than

peer-reviewed journal articles. More informative guidelines are likely to emerge with continued research, but many complex issues related to model identification, sample size, model complexity, and model fit are currently not clear cut. This study may raise awareness of the importance and implications of multilevel validity, and it is this increased attention to new issues that often spurs advancement of methodological capabilities.

CHAPTER 4

RESULTS

Using data from approximately 28,000 students in grades 3, 5, and 8 in a northeastern U.S. state, this study examined the multilevel factor structure of mathematics achievement tests. Multilevel exploratory factor analyses were used to determine the optimal number of factors at both the student and the school levels of analysis. Multilevel confirmatory factor analyses were used to assess the extent to which the one-factor solutions on each level were feasible. Both standard (single level) confirmatory factor analyses and multilevel confirmatory factor analyses were used to compare the size and relative importance of factor loadings at the different levels of analysis. The statistical significance of the school-level factor loadings provided evidence about the extent to which the mathematics achievement test items were effective for discriminating between schools.

Descriptive Results

The mathematics achievement tests contained 59 operational items at each grade level, consisting of multiple choice, short answer, and extended response items (see Table 3.3). The items were unique to each grade level; that is, for example, Item 1 for grade 3 and Item 1 for grade 5 were different, grade-appropriate items. Item means and standard deviations for each grade level are presented in Appendix A. For the results shown in Appendix A, as well as for all other results presented in this chapter, values have been rounded to two decimal places for the purpose of presentation in the tables.

The intraclass correlations (ICCs) for each item were produced by the multilevel exploratory and confirmatory factor analyses. The ICCs provide a descriptive measure of the proportion of school-level variation for each item. ICCs close to zero indicate that nearly all variation is at the student level, whereas ICCs close to 1.00 indicate that nearly all variation is at the school level. The typical range for ICCs in mathematics achievement is approximately .20 – .30 (Hedges & Hedberg, 2007).

The ICCs for each item by grade level are presented in Tables 4.1 – 4.3. The ICCs appeared to be smallest for grades 3 and 5, with many values around 0.10 or lower, and no values of 0.20 or higher. In contrast, the majority of items in grade 8 had ICCs above 0.10, and 16 of the 59 items had ICCs above 0.20. This pattern of ICCs by grade level suggests that more school-level variation in mathematics achievement exists by the end of middle school than during the elementary school years.

Table 4.1

Grade 3 Intraclass Correlations (ICCs), by Item

Item	ICC	Item	ICC	Item	ICC
1	.07	21	.09	41	.03
2	.05	22	.04	42	.07
3	.04	23	.07	43	.13
4	.11	24	.09	44	.11
5	.07	25	.14	45	.08
6	.08	26	.18	46	.08
7	.07	27	.11	47	.10
8	.10	28	.11	48	.05
9	.06	29	.11	49	.07
10	.13	30	.12	50	.09
11	.09	31	.13	51	.08
12	.11	32	.13	52	.09
13	.18	33	.05	53	.06
14	.09	34	.10	54	.08
15	.07	35	.11	55	.16
16	.14	36	.04	56	.11
17	.10	37	.04	57	.09
18	.04	38	.08	58	.09
19	.07	39	.15	59	.03
20	.06	40	.10		

Table 4.2

Grade 5 Intraclass Correlations (ICCs), by Item

Item	ICC	Item	ICC	Item	ICC
1	.03	21	.07	41	.01
2	.05	22	.05	42	.15
3	.04	23	.04	43	.05
4	.03	24	.14	44	.07
5	.10	25	.03	45	.04
6	.01	26	.05	46	.05
7	.08	27	.04	47	.04
8	.06	28	.11	48	.07
9	.02	29	.11	49	.03
10	.08	30	.09	50	.03
11	.05	31	.07	51	.10
12	.03	32	.03	52	.10
13	.07	33	.14	53	.08
14	.12	34	.09	54	.15
15	.09	35	.10	55	.03
16	.02	36	.10	56	.05
17	.08	37	.06	57	.07
18	.08	38	.05	58	.04
19	.10	39	.02	59	.01
20	.08	40	.05		

Table 4.3

Grade 8 Intraclass Correlations (ICCs), by Item

Item	ICC	Item	ICC	Item	ICC
1	.15	21	.13	41	.08
2	.11	22	.27	42	.22
3	.02	23	.07	43	.12
4	.22	24	.14	44	.16
5	.11	25	.14	45	.22
6	.25	26	.23	46	.06
7	.02	27	.16	47	.18
8	.09	28	.26	48	.14
9	.14	29	.26	49	.15
10	.15	30	.24	50	.14
11	.30	31	.29	51	.22
12	.32	32	.15	52	.19
13	.13	33	.12	53	.17
14	.29	34	.22	54	.11
15	.23	35	.18	55	.10
16	.08	36	.11	56	.12
17	.09	37	.11	57	.14
18	.16	38	.14	58	.07
19	.11	39	.06	59	.14
20	.06	40	.16		

Optimal Number of Factors at Each Level of Analysis

Research Question 1 concerned the optimal number of factors present at each level of analysis. This question was addressed using multilevel exploratory factor analyses. The eigenvalues, interpretability, and the fit criteria described in Table 3.4 were all used to arrive at the optimal number of factors for each level and grade. First, the eigenvalues at both the within and between levels of analysis were used to construct scree plots. Sample scree plots for each level of analysis at grade 3 are presented in Figures 4.1 and 4.2.

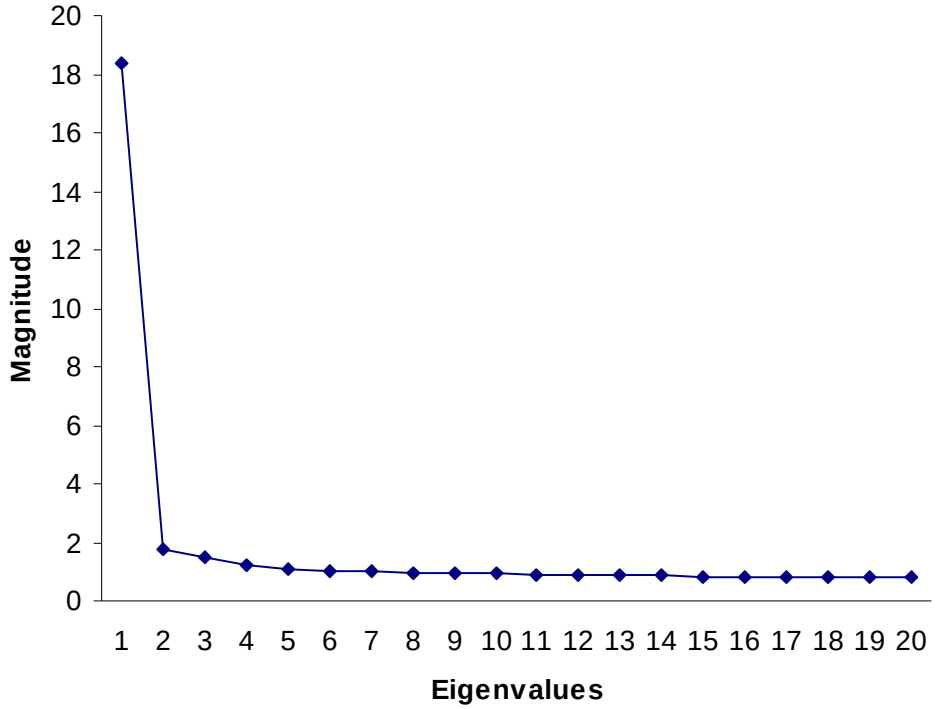

Figure 4.1. Scree Plot of First 20 Within-level Eigenvalues for Grade 3 Data

The pattern of eigenvalues was remarkably similar for each grade, with the only steep drop occurring after the first eigenvalue. This was true for both the student and school levels of analysis. The magnitude of the largest eigenvalues at the between level appeared much greater than the largest eigenvalues at the within level. The scree plots provide a rough

estimate of the number of factors present at each level of analysis, plus or minus a couple of factors. Based on the results from the scree plots, multilevel exploratory factor analyses were performed for 1-3 factors at each level of analysis. Models were also run where the factor structure was unrestricted (just-identified) at one level of analysis. When one level of analysis provides perfect fit (as occurs when the model is just-identified), this procedure facilitates the process of identifying a sufficient number of factors at the other level of analysis (Muthén & Asparouhov, 2009).

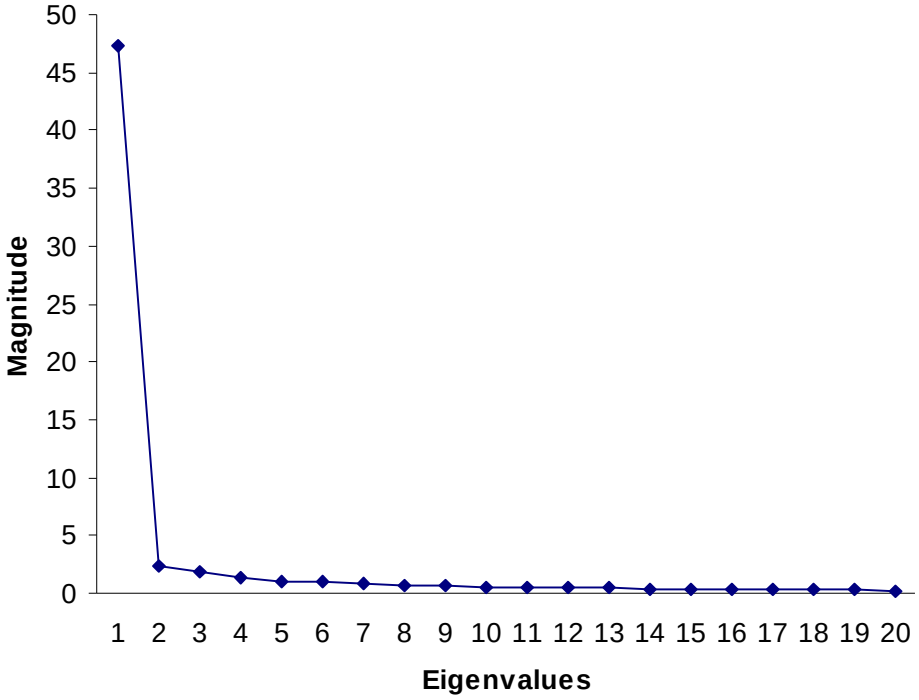

Figure 4.2. Scree Plot of First 20 Between-level Eigenvalues for Grade 3 Data

Multilevel exploratory factor analyses were conducted on all combinations of models with one to three factors on each level, along with unrestricted models for one level at a time. This yielded a total of 45 multilevel exploratory factor analysis solutions (15 at each grade level).

All 15 solutions at each grade level were considered and evaluated. Fit statistics for the most relevant subset of models are presented in Tables 4.4 – 4.6. Fit statistics for the full set of 45 models can be found in Appendix B.

Table 4.4

Two-level Exploratory Factor Analyses Fit Statistics for Grade 3

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	2,855.09* (1652)	1.00	1.00	.01	.00	.06
1	UN	7,876.13* (1652)	.99	.98	.02	.03	.00
1	1	14,374.00* (3304)	.98	.98	.02	.03	.06
2	2	10,145.22* (3188)	.99	.99	.02	.03	.05
3	3	7,099.61* (3074)	.99	.99	.01	.02	.04

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

Table 4.4 provides information about the fit of the multilevel exploratory factor analyses for grade 3. The use of an unrestricted model at one level can provide valuable information about the fit of the model at the other level of analysis. Here the unrestricted model at the within level produced adequate fit statistics for one factor at the between level. With the exception of the chi-square statistic, all other fit statistics for the one factor between model met the criteria established previously and listed in Table 3.4. It should be noted that none of the 15 models for grade 3 produced a nonsignificant chi-square statistic. Given the

large sample size of approximately 9,000 students in grade 3, it would be unlikely to find a parsimonious model with a nonsignificant chi-square statistic.

The use of an unrestricted model at the between level also suggested that a one factor solution at the within level may be sufficient. With the exception of the chi-square statistic, all other fit statistics met the recommended criteria. Likewise, the model with one factor at each level also produced adequate fit statistics with the exception of the chi-square statistic.

Although the one factor solution at each level appeared to provide adequate fit for grade 3, solutions with two and three factors resulted in slightly better fit. Although the improvement in most fit statistics is likely statistically significant, it is expected that additional factors will improve fit even when they are largely noise. The patterns of factor loadings for each combination of factors were examined for size and interpretability. In all cases, solutions with multiple factors resulted in one dominant factor, with few to no large loadings on the other factor(s). For example, in the model with two factors at each level (see Appendix C), nearly all items had moderate to strong loadings (.4 – .7) on the first factor at the within level, and only 1 of the 59 items had a loading of .3 or greater on the second factor at the within level. At the between level, all items had strong loadings (.6 – 1.0) on the first factor, and only 3 of the 59 items had loadings of .3 or greater on the second factor. This pattern was evident for all models with multiple factors at one or more levels. Although the majority of factor loadings were statistically significant, the secondary loadings did not account for much explained variation. These findings suggest that the one factor solution at each level provided the most parsimonious model for grade 3, while models with additional factors were evidence of overfactoring. The solution for one factor at each level of analysis is presented in Appendix D.

For grade 5, the results for the number of factors at the between level were mixed (see Table 4.5). When an unrestricted model was used at the within level, the one factor solution at the between level met the recommended criteria for the TLI, CFI, and RMSEA, but did not meet the standard for the SRMR at the between level or the chi-square statistic.

Table 4.5

Two-level Exploratory Factor Analyses Fit Statistics for Grade 5

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	2,255.57* (1652)	1.00	1.00	.01	.00	.10
1	UN	5,551.48* (1652)	.99	.98	.02	.03	.00
1	1	9,649.71* (3304)	.99	.99	.02	.03	.10
2	2	7,258.78* (3188)	.99	.99	.01	.02	.09
3	3	5,920.06* (3074)	.99	.99	.01	.02	.08

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

As mentioned previously, the chi-square statistic is very sensitive to sample size, and none of the 15 models run on the approximately 9,000 students in grade 5 produced a nonsignificant chi-square statistic. The SRMR at the between level, however, suggests that additional factors could improve model fit, as the models with three factors on the between level did meet the recommended guideline of .08 for the SRMR. At the within level, the models with

one factor did provide adequate fit, with the exception of the chi-square statistic. Models with additional factors provided slightly better fit.

When the factor loadings were examined for solutions with multiple factors at one or more levels, however, none of the solutions were interpretable. As with grade 3, there was strong evidence of one primary factor, with the additional factors having few or no moderate loadings. Consequently, the solution with one factor at each level was judged to be most parsimonious for grade 5. Factor loadings for this solution are provided in Appendix D.

Table 4.6

Two-level Exploratory Factor Analyses Fit Statistics for Grade 8

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	1,068.96 (1652)	1.00	1.00	.00	.00	.05
1	UN	9,204.67* (1652)	.98	.97	.02	.04	.00
1	1	16,200.54* (3304)	.97	.97	.02	.04	.05
2	2	5,372.34* (3188)	1.00	1.00	.01	.02	.04
3	3	4,015.77* (3074)	1.00	1.00	.01	.02	.03

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

For grade 8, the unrestricted model at the within level provided evidence that a single factor was sufficient at the between level (see Table 4.6). This solution met the criteria for all fit statistics, including the chi-square statistic. The unrestricted model at the between level,

with one factor at the within level, resulted in adequate fit with the exception of the chi-square statistic. It should be noted that the chi-square statistic was significant for all solutions where the within level was estimated, which is not surprising given the approximately 10,000 students included in the grade 8 analyses. Although the one factor solution at each level generally provided adequate fit, the fit did improve slightly with additional factors. However, as with the solutions for grades 3 and 5, the additional factors had few to no high loadings and were not interpretable. Therefore, the one factor solution was preferred as most parsimonious. Factor loadings for this model are also provided in Appendix D.

Feasibility of One Factor Solution at Each Level of Analysis

The second research question investigated the extent to which one factor solutions at each level were feasible, regardless of how many factors were optimal. Multilevel confirmatory factor analyses were performed for each grade, with a single factor at the student level and a single factor at the school level. Given the findings of Research Question 1, it appears that the one factor solutions were not only feasible but were the most interpretable and parsimonious. The fit statistics were the same as those for the one factor multilevel exploratory factor analyses (see Tables 4.4 – 4.6) and were previously discussed in regard to the first research question. Standardized factor loadings and standard errors from the one factor multilevel confirmatory factor analyses are presented on the left side of Tables 4.7 – 4.9. Standardized factor loadings and standard errors from the single level confirmatory factor analyses (where the student and school levels are collapsed) are presented on the right side of the same tables for purposes of comparison. Relevant findings from the collapsed solutions will be discussed in regard to Research Question 3.

Factor loadings at the within and between levels were examined as indicators of item quality. A loading that was near zero or negative would indicate that an item was not contributing much to (or was even detracting from) the measurement of the construct at a given level of analysis. For grade 3, the majority of the within-level loadings were moderate, ranging from .26 – .73. None of the within-level loadings for grade 3 were negative or close to zero. The between-level loadings were uniformly strong for grade 3, ranging from .72 – 99. Consequently, for grade 3, it appears that all items contributed to the measurement of the construct at both the student and school levels of analysis.

Table 4.7

Confirmatory Factor Analyses Solutions for Grade 3

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.58* (.01)	.85* (.04)	.60* (.01)
2	.26* (.02)	.79* (.08)	.31* (.02)
3	.34* (.01)	.78* (.06)	.37* (.01)
4	.65* (.01)	.94* (.02)	.69* (.01)
5	.61* (.01)	.84* (.04)	.62* (.01)
6	.66* (.01)	.95* (.02)	.68* (.01)
7	.47* (.01)	.80* (.04)	.50* (.01)
8	.61* (.01)	.91* (.03)	.64* (.01)
9	.50* (.01)	.95* (.03)	.53* (.01)
10	.57* (.01)	.80* (.04)	.60* (.01)
11	.46* (.01)	.87* (.03)	.50* (.01)
12	.63* (.01)	.93* (.02)	.67* (.01)
13	.71* (.01)	.92* (.02)	.74* (.01)
14	.58* (.01)	.99* (.01)	.63* (.01)
15	.46* (.01)	.91* (.03)	.51* (.01)
16	.72* (.01)	.95* (.02)	.75* (.01)
17	.58* (.01)	.84* (.04)	.61* (.01)
18	.46* (.01)	.96* (.03)	.48* (.01)
19	.62* (.01)	.96* (.02)	.64* (.01)
20	.63* (.01)	.95* (.02)	.65* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.7 (continued)

Confirmatory Factor Analyses Solutions for Grade 3

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.54* (.01)	.87* (.03)	.57* (.01)
22	.26* (.01)	.87* (.04)	.31* (.01)
23	.54* (.01)	.83* (.04)	.56* (.01)
24	.53* (.01)	.94* (.02)	.57* (.01)
25	.46* (.02)	.91* (.03)	.53* (.01)
26	.64* (.01)	.92* (.02)	.69* (.01)
27	.65* (.01)	.91* (.02)	.68* (.01)
28	.59* (.01)	.90* (.03)	.63* (.01)
29	.64* (.01)	.92* (.02)	.67* (.01)
30	.44* (.01)	.80* (.04)	.49* (.01)
31	.54* (.01)	.92* (.02)	.60* (.01)
32	.59* (.01)	.96* (.01)	.64* (.01)
33	.39* (.01)	.94* (.03)	.43* (.01)
34	.53* (.01)	.91* (.03)	.58* (.01)
35	.73* (.01)	.91* (.03)	.76* (.01)
36	.32* (.02)	.73* (.09)	.35* (.02)
37	.46* (.02)	.83* (.06)	.48* (.02)
38	.65* (.01)	.98* (.02)	.68* (.01)
39	.55* (.01)	.80* (.04)	.59* (.01)
40	.56* (.01)	.91* (.03)	.60* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.7 (continued)

Confirmatory Factor Analyses Solutions for Grade 3

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.43* (.01)	.91* (.04)	.45* (.01)
42	.52* (.01)	.92* (.03)	.55* (.01)
43	.61* (.01)	.76* (.05)	.63* (.01)
44	.58* (.01)	.72* (.06)	.60* (.01)
45	.61* (.01)	.98* (.02)	.65* (.01)
46	.69* (.01)	.94* (.02)	.71* (.01)
47	.65* (.01)	.90* (.03)	.67* (.01)
48	.50* (.02)	.96* (.07)	.53* (.02)
49	.61* (.01)	.98* (.02)	.64* (.01)
50	.56* (.01)	.96* (.02)	.60* (.01)
51	.50* (.01)	.88* (.04)	.54* (.01)
52	.48* (.01)	.91* (.03)	.53* (.01)
53	.34* (.02)	.93* (.05)	.39* (.02)
54	.41* (.01)	.81* (.04)	.45* (.01)
55	.46* (.02)	.63* (.07)	.50* (.02)
56	.53* (.01)	.91* (.03)	.57* (.01)
57	.48* (.01)	.87* (.03)	.52* (.01)
58	.65* (.01)	.97* (.02)	.69* (.01)
59	.50* (.01)	.96* (.05)	.50* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.8

Confirmatory Factor Analyses Solutions for Grade 5

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.55* (.01)	.96* (.03)	.57* (.01)
2	.49* (.01)	.98* (.03)	.52* (.01)
3	.47* (.01)	.96* (.03)	.49* (.01)
4	.41* (.01)	.94* (.04)	.44* (.01)
5	.48* (.01)	.81* (.05)	.51* (.01)
6	-.02 (.02)	-.13 (.20)	-.02 (.02)
7	.62* (.01)	.93* (.03)	.64* (.01)
8	.56* (.01)	.99* (.02)	.59* (.01)
9	.36* (.01)	.86* (.07)	.38* (.01)
10	.45* (.01)	.77* (.05)	.47* (.01)
11	.56* (.01)	.94* (.02)	.58* (.01)
12	.36* (.01)	.86* (.05)	.38* (.01)
13	.62* (.01)	.93* (.02)	.64* (.01)
14	.54* (.01)	.71* (.06)	.56* (.01)
15	.61* (.01)	.91* (.03)	.63* (.01)
16	.30* (.01)	.98* (.03)	.33* (.01)
17	.63* (.01)	.97* (.02)	.65* (.01)
18	.44* (.01)	.54* (.09)	.45* (.01)
19	.58* (.01)	.90* (.03)	.62* (.01)
20	.61* (.01)	.90* (.03)	.63* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.8 (continued)

Confirmatory Factor Analyses Solutions for Grade 5

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.59* (.01)	.88* (.03)	.61* (.01)
22	.56* (.01)	.91* (.04)	.58* (.01)
23	.51* (.01)	.93* (.03)	.53* (.01)
24	.39* (.01)	.56* (.07)	.41* (.01)
25	.46* (.01)	.96* (.03)	.48* (.01)
26	.50* (.01)	.80* (.05)	.51* (.01)
27	.39* (.01)	.76* (.07)	.40* (.01)
28	.62* (.01)	.72* (.05)	.63* (.01)
29	.55* (.01)	.94* (.02)	.58* (.01)
30	.55* (.01)	.97* (.01)	.58* (.01)
31	.59* (.01)	.93* (.03)	.62* (.01)
32	.37* (.01)	.71* (.08)	.38* (.01)
33	.46* (.01)	.24* (.11)	.43* (.01)
34	.52* (.01)	.94* (.03)	.56* (.01)
35	.50* (.01)	.59* (.08)	.50* (.01)
36	.63* (.01)	.91* (.03)	.66* (.01)
37	.51* (.01)	.89* (.04)	.54* (.01)
38	.43* (.02)	.82* (.06)	.46* (.01)
39	.28* (.02)	.77* (.11)	.29* (.02)
40	.41* (.01)	.77* (.06)	.43* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.8 (continued)

Confirmatory Factor Analyses Solutions for Grade 5

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.14* (.02)	.10 (.17)	.13* (.02)
42	.43* (.01)	.53* (.08)	.44* (.01)
43	.59* (.01)	.83* (.05)	.60* (.01)
44	.56* (.01)	.91* (.03)	.59* (.01)
45	.64* (.01)	.98* (.03)	.66* (.01)
46	.57* (.01)	.94* (.03)	.59* (.01)
47	.49* (.01)	.82* (.06)	.50* (.01)
48	.39* (.02)	.76* (.06)	.41* (.01)
49	.38* (.02)	.88* (.06)	.40* (.02)
50	.52* (.02)	.94* (.09)	.54* (.02)
51	.47* (.01)	.84* (.04)	.51* (.01)
52	.62* (.01)	.96* (.02)	.65* (.01)
53	.49* (.01)	.72* (.06)	.51* (.01)
54	.53* (.01)	.82* (.04)	.56* (.01)
55	.40* (.02)	.95* (.06)	.42* (.02)
56	.28* (.02)	.73* (.08)	.31* (.01)
57	.56* (.01)	.96* (.03)	.59* (.01)
58	.52* (.01)	.95* (.03)	.55* (.01)
59	.15* (.01)	.68* (.19)	.16* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.9

Confirmatory Factor Analyses Solutions for Grade 8

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.51* (.02)	.94* (.02)	.55* (.01)
2	.49* (.01)	.93* (.04)	.53* (.01)
3	.21* (.01)	.47* (.10)	.20* (.01)
4	.56* (.01)	.92* (.03)	.60* (.01)
5	.43* (.02)	.88* (.03)	.45* (.01)
6	.40* (.01)	.83* (.06)	.47* (.01)
7	.05* (.01)	-.29* (.12)	.02 (.01)
8	.34* (.01)	.69* (.08)	.39* (.02)
9	.51* (.01)	.88* (.04)	.56* (.01)
10	.57* (.02)	.87* (.05)	.58* (.01)
11	.67* (.01)	.96* (.05)	.73* (.01)
12	.70* (.02)	.95* (.08)	.72* (.01)
13	.56* (.01)	.97* (.02)	.60* (.01)
14	.73* (.01)	.96* (.06)	.76* (.01)
15	.64* (.01)	.93* (.06)	.68* (.01)
16	.45* (.01)	.97* (.02)	.48* (.01)
17	.47* (.01)	.97* (.02)	.50* (.01)
18	.55* (.01)	.99* (.01)	.59* (.01)
19	.50* (.01)	.98* (.01)	.53* (.01)
20	.43* (.01)	.88* (.04)	.44* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.9 (continued)

Confirmatory Factor Analyses Solutions for Grade 8

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.56* (.01)	.96* (.02)	.60* (.01)
22	.72* (.01)	.96* (.05)	.76* (.01)
23	.44* (.01)	.96* (.02)	.47* (.01)
24	.50* (.01)	.98* (.01)	.55* (.01)
25	.61* (.01)	.98* (.01)	.66* (.01)
26	.56* (.01)	.93* (.03)	.61* (.01)
27	.50* (.01)	.95* (.03)	.55* (.01)
28	.61* (.01)	.98* (.03)	.68* (.01)
29	.72* (.01)	.96* (.02)	.76* (.01)
30	.61* (.01)	.94* (.02)	.66* (.01)
31	.72* (.01)	.95* (.05)	.74* (.01)
32	.54* (.01)	.90* (.03)	.56* (.01)
33	.56* (.01)	1.00* (.01)	.59* (.01)
34	.67* (.02)	.71* (.08)	.66* (.01)
35	.68* (.01)	.98* (.02)	.72* (.01)
36	.52* (.01)	.89* (.03)	.56* (.01)
37	.48* (.01)	.96* (.02)	.53* (.01)
38	.55* (.01)	.96* (.02)	.59* (.01)
39	.23* (.02)	.68* (.07)	.26* (.01)
40	.64* (.01)	.97* (.02)	.67* (.01)

Note. SE = Standard Error.
* $p < .05$

Table 4.9 (continued)

Confirmatory Factor Analyses Solutions for Grade 8

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.19* (.02)	.56* (.09)	.22* (.01)
42	.71* (.01)	.99* (.02)	.75* (.01)
43	.55* (.01)	.94* (.02)	.61* (.01)
44	.51* (.01)	.98* (.01)	.55* (.01)
45	.61* (.01)	.93* (.02)	.65* (.01)
46	.39* (.01)	.78* (.07)	.39* (.01)
47	.61* (.01)	.94* (.04)	.63* (.01)
48	.52* (.01)	.96* (.01)	.56* (.01)
49	.49* (.01)	.96* (.02)	.54* (.01)
50	.46* (.01)	.89* (.03)	.50* (.01)
51	.61* (.01)	.94* (.02)	.65* (.01)
52	.63* (.01)	.91* (.03)	.66* (.01)
53	.57* (.01)	.95* (.02)	.61* (.01)
54	.54* (.01)	.99* (.02)	.57* (.01)
55	.45* (.01)	.97* (.02)	.49* (.01)
56	.52* (.01)	.97* (.01)	.55* (.01)
57	.48* (.01)	.99* (.01)	.53* (.01)
58	.38* (.01)	.91* (.05)	.41* (.01)
59	.50* (.01)	.97* (.01)	.56* (.01)

Note. SE = Standard Error.
* $p < .05$

For grade 5, there was one item at the within level (Item 6) that appeared not to be contributing anything to the measurement of student achievement, with a factor loading near zero (-.02). Items 41 and 59 contributed little at the within level, with factor loadings of .14 and .15, respectively. The remainder of the within-level loadings were moderate, ranging from .28 – .63. At the between level, Items 6 and 41 appeared not to be contributing to the measurement of school achievement, with factor loadings of -.13 and .10, respectively. The factor loading for Item 33 was also fairly low at .24. The remainder of the between-level loadings were moderate to strong, ranging from .53 – .98.

For grade 8, one item at the within level (Item 7) had a factor loading of .05, implying that this item was not a very good indicator of student achievement. Items 3, 39, and 41 appeared to be only weakly related to the measurement of student achievement, with loadings of approximately .20. The remainder of the within-level loadings were moderate, ranging from .34 – .73. At the between level, there was only one item (Item 7) that did not appear to be a strong indicator of school achievement. In fact, with a loading of -.29, this item actually detracted from the measurement of school achievement. All other loadings at the between level were moderate to high, with the majority close to .9 or above.

Comparison of Factor Loadings at Different Levels of Analysis

The third research question involved comparisons of the factor loadings at the within and between levels of analysis, and to standard (single level) factor loadings where the multilevel data structure was collapsed (i.e., ignored). First, the within- and between-level factor loadings were compared descriptively. For all grades, the between-level factor loadings generally appeared to be much larger than the within-level factor loadings (see Tables 4.7 – 4.9). The majority of the between-level factor loadings were .7 or above,

whereas the majority of the within-level factor loadings were well below .7. This finding implies that in most cases, items appeared much *more* discriminating at the school level than at the student level.

There were a few notable exceptions to this pattern. In grade 5 (see Table 4.8), Item 33 appeared to have a smaller loading at the between level (.24) than at the within level (.46). Items 6 and 41 had low loadings at both the within and between levels. In grade 8 (see Table 4.9), Item 7 appeared to have a smaller loading at the between level (-.29) than at the within level (.05). Item 34 had high loadings at both the within and between levels (.67 and .71, respectively).

In addition, the relative standing of item factor loadings was not necessarily uniform across levels. Items with the highest loadings at one level did not necessarily display the highest loadings at the other level. For example, in grade 8 (see Table 4.9), Item 33 had the highest between-level factor loading of 1.00; the within-level loading for this item was .56, which fell in the middle of the range of values for the student-level factor loadings. This pattern was evident across all grades. Consequently, the descriptive comparisons of factor loadings across the student and school levels suggested that items differed in terms of both their absolute and relative standings.

The second step in comparing the within- and between-level factor loadings was to conduct scaled chi-square difference tests. Models where the loadings were freely estimated at each level were compared to nested models where factor loadings were constrained to be equal across levels. The results of the scaled chi-square difference tests are presented in Table 4.10.

For all grades, the more constrained model (where factor loadings were equal across levels) resulted in significantly worse fit. This finding indicates that the magnitude of the factor loadings was not the same for the student and school levels of analysis.

Table 4.10

Chi-square Difference Tests, by Grade

Grade	$\chi^2_{difference}$	$df_{difference}$
3	293.49*	59
5	150.88*	59
8	233.36*	59

* $p < .05$

Finally, standard confirmatory factor analyses were performed using a single level of analysis for each grade. The factor loadings and standard errors are included on the right side of Tables 4.7 – 4.9. Although there are no statistical tests available for comparing the magnitude of the standard factor loadings to the multilevel factor loadings, the descriptive comparisons are informative. For all grades, the standard factor loadings were very similar to the within-level factor loadings, generally within .05. The between-level factor loadings were almost always higher than the standard factor loadings. For example, in grade 3 (see Table 4.7), the collapsed loading for Item 1 was .60, which is very close to the within-level estimate of .58; this appears to be much lower than the between-level estimate of .85.

The similarity of the standard and within-level factor loadings also applied to items that did not appear to be strongly related to the overall construct at one or more levels. For example, in grade 5 (see Table 4.8), the collapsed factor loading for Item 6 was -.02, the same value as the within-level estimate; the between-level estimate, also near zero, was -.13. The collapsed factor loading for Item 33 was .43, very similar to the within-level estimate of .46; the between-level estimate for this item appeared to be lower at .24.

The comparisons of the single level loadings to the multilevel loadings suggest that the process of performing traditional factor analyses (i.e., ignoring the student and school levels) provided reasonable approximations for the within-level loadings but yielded no information about the absolute or relative size of the between-level loadings. The between-level factor loadings were generally larger, occasionally smaller, and universally different than the factor loadings obtained in analyses that ignored the multilevel structure by performing standard factor analyses on the collapsed data.

Statistical Significance of Between-level Factor Loadings

The final research question concerned the statistical significance of the factor loadings, specifically those at the school level. Significance tests were performed by dividing each factor loading by its standard error; resulting values greater than 1.96 were statistically significant at $p < .05$. The statistical significance of the factor loadings is noted in Tables 4.7 – 4.9. Nearly all factor loadings were statistically significant, across grades and levels of analysis. Only a small number of between-level factor loadings were nonsignificant. In addition to being statistically significant, the factor loadings appeared to be meaningful, as the majority were approximately .9 or higher. This finding indicates that all of the observed

variance in the school-level loadings was explained by the common factor of school mathematics achievement.

In grade 3 (see Table 4.7), all between-level factor loadings were positive and statistically significant. There was not a single school-level factor loading that was nonsignificant, and the smallest between-level loading was .63. This suggests that all items in grade 3 made strong contributions to the measurement of school achievement.

In grade 5 (see Table 4.8), two school-level factor loadings were statistically nonsignificant (i.e., not different from zero); Item 6 had a between-level factor loading of -.13 and Item 41 had a between-level factor loading of .10. This finding suggests that Items 6 and 41 for grade 5 did not contribute to the measurement of school-level achievement. The other 57 school-level factor loadings were statistically significant. Although two items did not appear to contribute to the measurement of school mathematics achievement in grade 5, the majority of the items did appear to make strong contributions.

In grade 8 (see Table 4.9), all between-level factor loadings were statistically significant, but the factor loading for Item 7 was negative. Although the magnitude of this factor loading was relatively small (-.29), the negative value indicates that Item 7 actually detracted from the measurement of school-level mathematics achievement. All other between-level factor loadings for grade 8 were positive and most were very strong. Overall, the collection of items that constituted the grade 8 test made strong contributions to the measurement of school achievement in mathematics.

Alternative Method of Model Identification

To investigate whether the method of model identification (setting the variance to one at each level of analysis) may have artificially inflated any effects across levels due to differences in variances, the multilevel confirmatory factor analysis was re-run for grade 3. The first factor loading was set to one on each level of analysis, which put all factor loadings on the scale of the Item 1. Results of the unstandardized factor loadings with this method of model identification are presented in Appendix E.

With this alternative approach to model identification, the discrepancies between the within-level factor loadings and the between-level factor loadings do appear to be somewhat smaller for the unstandardized solution, and the standard errors appear larger. The difference in the magnitude of the factor loadings across levels is difficult to interpret given the difference in scaling. In nearly all cases, the between-level factor loadings appear to remain larger than the within-level factor loadings. The interpretation of this finding is unclear, given that the factor loadings from this solution are not on the same scale at each level. That is, the within-level factor loadings are relative to the size of the within-level loading for Item 1, and the between-level factor loadings are relative to the size of the between-level loading for Item 1. It also is not clear how results might vary depending on the characteristics of the item that is used to identify the scale. For example, Van den berg, Glas, and Boomsma (2007) caution against this method of model identification when the measurement model is of interest because the standard errors may be affected by characteristics of the item set to unity. Despite the potential problem caused by implicitly assuming that the variance is the same at each level of analysis, both the absolute and relative analyses undertaken here appear

more meaningful when the model is identified by setting the variance to one instead of the first factor loading.

Follow-up Analyses Excluding Small Clusters

There was a concern that the presence of small clusters (schools with few students at a particular grade level) could bias the parameters and standard errors. To assess the potential impact of very small clusters, all confirmatory factor analyses were re-run using only those schools with five or more students at a given grade. Descriptive statistics on the number of schools and number of students per school at each grade level for the follow-up analyses are presented in Table 4.11. The follow-up analyses excluded data from five schools at grade 3, nine schools at grade 5, and eight schools at grade 8.

Results from the follow-up analyses were very similar to the original analyses; all factor loadings from the follow-up analyses are presented in Appendix F. The majority of the factor loadings were the same (to two decimal places), several factor loadings differed by .01, and a small number of factor loadings differed by .02. In no case did the factor loadings differ by more than .02, and the statistical significance of the factor loadings was unchanged. This finding indicates that it is unlikely that the inclusion of small clusters (with less than five students per school at a given grade) affected the results in this study.

Table 4.11

Number of Students per School by Grade (Limited to Clusters of Five or More)

Test	Schools	Number of Students per School			
		Minimum	Maximum	Mean	SD
Mathematics Grade 3	106	5	180	84.70	36.68
Mathematics Grade 5	87	5	395	102.97	75.70
Mathematics Grade 8	76	5	541	134.43	147.75

Summary

Four research questions about the validity of school-level inferences were investigated using mathematics achievement data from all students in grades 3, 5, and 8 who participated in a state-wide mathematics achievement test. The first research question investigated the optimal number of factors at each level of analysis. For all three grades, one factor was found to be optimal at both the student and school levels of analysis. Although solutions with two or three factors resulted in slightly better fit statistics, the solutions were not interpretable or meaningful because the additional factors contained few or no items with moderate or high loadings. The solutions with one factor at each level were the most parsimonious at each grade.

The second research question addressed the feasibility of the one factor solutions at each level, regardless of how many factors were optimal. Given the findings of the first research question, the one factor solutions were judged to be both feasible and optimal.

Factor loadings at both the student and school levels were examined as indicators of item quality. Across all three grades, the majority of factor loadings were moderate at the within level and strong at the between level. In grade 5, three factor loadings at the within level were near zero or very low (.2 or below), while two factor loadings at the between level were near zero. In grade 8, four factor loadings at the within level were near zero or very low (.2 or below), while one factor loading at the between level was negative. Overall, most items appeared to be contributing to the measurement of student mathematics achievement and school mathematics achievement, and this finding was applicable to all grades.

The third research question involved comparisons of factor loadings across different levels of analysis. The within- and between-level factor loadings were compared both descriptively and with chi-square difference tests. Both the relative and absolute size of the loadings were noted across levels. Across all grades, the majority of the within-level factor loadings were moderate (.4 – .7), while the majority of the between-level factor loadings were strong (.7 or above). This finding implies that most items were more discriminating at the school level than at the student level. In addition, the relative standing of item factor loadings was not the same across levels. For all grades, the more constrained model (where factor loadings were equal across levels) resulted in significantly worse fit, indicating that the magnitude of the factor loadings was not the same for the student and school levels of analysis.

Research Question 3 also involved comparisons of factor loadings from standard (single level) confirmatory factor analyses to the within- and between-level factor loadings. For all grades, the standard factor loadings were close to the within-level factor loadings (nearly always within .05) but were generally not close to the between-level factor loadings.

The between-level factor loadings were generally larger, occasionally smaller, and often very different than the factor loadings obtained in standard factor analyses that ignored the multilevel structure. This finding indicates that the process of performing traditional factor analyses in this study provided reasonable approximations for the within-level factor loadings but yielded no information about the absolute or relative size of the between-level factor loadings.

The final research question concerned the statistical significance of the factor loadings at the school level. Nearly all factor loadings were statistically significant, across grades and levels of analysis. In grade 3, all between-level factor loadings were statistically significant and positive. In grade 5, two between-level factor loadings were statistically nonsignificant; this finding indicates that these two items did not contribute to the measurement of school achievement. In grade 8, all between-level factor loadings were statistically significant, but the factor loading for one item was negative; this finding indicates that one item detracted from the measurement of school achievement. Overall, nearly all items at each grade level appeared to have statistically significant and strong contributions to the measurement of school achievement.

Follow-up analyses (limited to schools with five or more students at a given grade) produced results that were nearly identical to the full sample of students and schools. This finding indicates that the inclusion of small clusters in this study does not appear to have had a significant impact on the results.

The next chapter discusses the interpretation of these findings in the context of relevant research on validity and student achievement. Implications for educational measurement and program evaluation and ideas for future research are also addressed.

CHAPTER 5

DISCUSSION

It has been nearly 50 years since Ebel (1961) said of validity, “It is universally praised, but the good works done in its name are remarkably few” (p. 640). Validation is the most important aspect of the measurement process, and psychometric theory is clear about the need for collecting validity evidence for each intended test purpose. But despite the widespread use of student achievement tests for making school-level inferences, the psychometric literature is devoid of studies investigating the adequacy of validity evidence for this purpose. Recent studies of group-level psychometrics in other fields have found that validity evidence is not necessarily uniform across multiple levels of analysis. This study extends the emerging body of research on multilevel construct validation to school-level achievement as currently encountered in K-12 student testing programs.

This final chapter begins with an acknowledgement of study limitations, followed by a discussion of the study findings. Study implications and areas worthy of future research are described. In addition to the specific research findings in the study, implications of bringing attention to the issue of multilevel validity in general are considered. The chapter concludes with recommendations for future directions in multilevel psychometrics.

Limitations

Before discussing the conclusions and implications of this research, it is important to acknowledge some limitations of the study sample, design, and analysis. First, the students

and state mathematics achievement tests were from a single state. It is not clear whether the students or the state testing program would be representative of those found in other states. Although the level one sample size of approximately 28,000 students is relatively large, the level two sample size (schools) ranged from 84 to 111—a relatively small sample for the analyses that were undertaken. It would be helpful to replicate the study using data from a much larger or more diverse state, such as Texas or California, to determine what aspects of the analyses or conclusions might be affected by sample size.

Second, although three different grades were examined, the study design included only one subject area, mathematics. It is not clear whether any of the results found here would apply to state achievement tests in other subjects. It would be helpful to replicate the study with data from state reading and science achievement tests to determine what aspects of the study results might be affected by subject area.

It is important to consider the limitations of study sample and design in the context of the study purpose. With data from only one state and subject area, this study was not intended to affirm or condemn the widespread practice of drawing school-level inferences from student level data. Instead, this exploratory study sought to illustrate how the question of looking at validity evidence for group-level inferences when using individual level data could be approached. A primary goal was to serve as a prototype for how such questions could be considered in future research.

Several analytic limitations were described previously in Chapter 3, most notably the limitation to two levels of analysis (student and schools), the assumption of homogeneity of the within-groups covariance matrix, and the novelty of methodological procedures for fitting multilevel factor analysis models with categorical variables. This study was constrained by

the methodology that is currently available, but it also serves as an opportunity to spur the additional research necessary for advancing the methodology in these (and other) areas. Methodological advances do not happen in a vacuum; there is an iterative process of asking new questions and improving the technology necessary to answer those questions. It is novel ideas in research and practice that serve as the impetus for improving methodology, and the methodological advances then generate additional questions. This study has the potential to increase the methodological capabilities in the area of multilevel validity so that future research is not bound by the same limitations faced here.

Key Findings

Overall, the present study yielded three key findings and implications. They include:

1) For each of the three grades studied, there was only one meaningful factor identified (presumably mathematics achievement) at both the student and school levels of analysis, providing tentative support for the current practice of drawing school-level inferences from student-level measures.

2) At each grade level, items differed in terms of both their absolute and relative size of their factor loadings at the student and school levels of analysis, suggesting that when school-level inferences are of interest, standard factor analyses provide insufficient information about test development and validation; when both within and between levels are of interest, factor loadings should be estimated separately.

3) The majority of items in this study were more discriminating at the school level than at the student level. Thus, if school-level inferences are of primary interest, it may be desirable to reconsider typical test construction practices in which items that fail to

discriminate at the student level (but unknowingly have adequate properties at the school level) are routinely removed from consideration for use on operational test forms.

Each of these findings is described in detail and implications of these findings are presented in the following sections.

Optimal Number of Factors at Each Level of Analysis

The first research question addressed in this research explored the optimal number of factors at the student and school levels using multilevel exploratory factor analysis. For all three grades studied, there appeared to be only one meaningful factor at both the student and school levels of analysis. Although solutions with additional factors resulted in slightly better fit, there were few to no high factor loadings on the additional factors. This finding indicates that there was only one primary factor at the student level (presumably student mathematics achievement) and one primary factor at the school level (presumably school mathematics achievement) that accounted for the underlying relationships between the items at each grade level.

The finding of one primary factor at each level of analysis is not surprising. The state mathematics achievement tests were designed to be single measures of student achievement, an assumption necessary for employing the unidimensional item response theory models used to develop, scale, and equate the tests. Given previous research on multilevel factor analysis, it would have been unlikely to uncover more factors at the school level than at the student level. Muthén (1989) found that the number of factors at the within level is an upper bound for the number of factors at the between level. Many studies using multilevel factor analysis have found the same number of factors at each level of analysis, and those that differ

have extracted fewer factors at the between level than the within level (Härnqvist et al, 1994; Hox, 2002; Kuhlemeier et al, 2002).

In this study, the consideration of the number of factors was not the primary focus of the research, but rather it was a prerequisite for performing subsequent analyses to compare factor loadings at different levels of analysis. If a single factor solution did not appear to fit the data adequately at one or more levels of analysis, then it would not be clear how to interpret potential differences in factor loadings across levels. A finding of different factor structures at the student and school level would itself indicate a threat to the validity of school-level inferences from student achievement data, however.

A more interesting question about the number of factors at each level of analysis could be investigated by examining several different achievement tests for the same students. For example, if mathematics, reading, and science data were analyzed for the same group of students, it is possible that there could be three separate achievement factors at the student level but only one general achievement factor at the school level. Such a finding would be consistent with Hox's (2002) analysis of verbal and numerical ability for children within families, where two separate factors were found at the student level but only a single general ability factor could be extracted at the family level. Future research in this area could reveal such differences across subject areas at the student and school levels of analysis. Failure to differentiate among subject areas at the school level could threaten the validity of using student achievement data to make school-level inferences about something that is ostensibly subject-specific if, in fact, the data were discovered to be more supportive of general achievement interpretations.

Evaluation of Factor Loadings at Multiple Levels of Analysis

Research Questions 2 through 4 concerned the size of the between-level factor loadings and the extent to which they differed from those at the within level and those obtained from traditional factor analyses where the multilevel structure was ignored (i.e., collapsed). The comparison of factor loadings at different levels of analysis involved both descriptive measures and scaled chi-square difference tests. At each grade level, the descriptive analyses found that items differed in terms of both their absolute and relative size of their factor loadings at the student and school levels of analysis. This result was confirmed by the scaled chi-square difference tests, which found that models where factor loadings were constrained to be equal across levels fit significantly worse than models where factor loadings were freely estimated at each level.

This particular finding of differences in the student and school level factor loadings has a couple of different implications. First, when factor loadings at the within and between levels are both of interest, it is necessary to estimate them separately. Neither the absolute or relative size of factor loadings at the within level provided much information about the factor loadings at the between level in this study. This suggests that the different sources of variation affected the factor loadings of each level in different ways. Second, even when the size of the factor loadings at the within and between levels are not of primary interest, it does not seem advisable to constrain the factor loadings to be equal across levels in any multilevel structural equation model of school achievement, despite the use of this practice for the purpose of reducing model parameters. This caution to avoid setting factor loadings equal at different levels of analysis is echoed by Muthén (2008), due to the fact that the item parameters have different meanings at each level.

The standard (single level) confirmatory factor analyses yielded factor loadings that were very similar to the within-level loadings in the multilevel confirmatory factor analysis; in nearly all cases, these two sets of factor loadings differed by less than .05. Conversely, the standard confirmatory factor analyses yielded no information about the relative or absolute size of the between-level loadings. Although accounting for the multilevel structure of students in schools is technically more accurate for both student- and school-level inferences, the implications of ignoring the multilevel structure appear much greater in the latter case. In this study, the use of standard factor analyses for test development and validation was a reasonable approximation for the student level results but provided very different information than the school level results. This indicates that when school-level inferences are of interest, standard factor analyses provide insufficient information about test development and validation. The school-level factor structure can only be obtained through multilevel confirmatory factor analyses; unlike the student-level factor structure, it cannot be reasonably approximated when the multilevel structure is ignored in traditional factor analyses that collapse both levels.

At each grade level, the vast majority of items had much higher factor loadings at the school level than at the student level. This finding was somewhat unexpected, particularly given the low amount of between-level variation reflected in the intraclass correlations. Because the factor loadings are standardized, it may be that the small amount of variation at the between level is largely accounted for by each item. It is not clear whether this result is typical in student achievement studies given the lack of similar research in the field.

The comparisons of factor loadings across levels were bound by the available options for model identification and procedures for standardization. The analyses undertaken here

have some parallels to studies of measurement invariance (with comparisons across levels instead of groups); the literature on measurement invariance generally recommends identifying the model by setting the first factor loading to one. The evaluation of item characteristics at each level of analysis draws on principles of item response theory, where model identification is achieved by setting the variance to one. Given the objectives of this study, the latter approach appeared to yield more meaningful and interpretable results; however, the impact of actual differences between student- and school-level variances is not clear. Guidelines for model identification and standardization procedures have not been developed in the context of multilevel validity. Future research in this area is needed, including the exploration of additional approaches for model identification and standardization in a multilevel context. For example, the model could be identified by setting the variance to one at the within level and setting the first factor loading to one at the between level.

The only other study known to have investigated item-level characteristics at two levels of analysis was performed by Reise et al. (2006) in the field of health care, who concluded that survey items were much less discriminating at the between (health plan) level than the within (individual) level. However, the approach taken by Reise at al. (2006) was quite different from this study; a three parameter multilevel item response theory model was used. The c-parameters were high at the between level (often .2 – .3) and this lower asymptote does not have an analog in the multilevel confirmatory factor analysis model with categorical variables. The survey used in the Reise et al. (2006) study had a ceiling effect, and responses on 3-4 point scales were recoded to be dichotomous. It is not clear how the

attenuation issues may have affected the item parameters at each level of analysis. Future research in this area is needed.

The most common explanation for obtaining higher factor loadings at the between level than the within level is that the aggregation process results in less error (Snijders & Bosker, 1999; Stanat & Lüdtke, 2008). Although it is often the case, the between-level reliability is not always higher than within-level reliability (Luppescu, Gladden, & Bryk, 2003). Unlike the within-level reliability, the between-level reliability is affected by both cluster size and between-cluster variability (Tate & King, 1994). In this study, the between-level factor loadings were not universally higher than the within-level factor loadings, even though this occurred most of the time.

Although the higher factor loadings found at the between level may be in part due to lower measurement error from the process of aggregation, psychometricians caution against assuming that school-level reliability is necessarily higher than student-level reliability. For example, Feldt and Brennan (1989) have noted that:

> [T]raditional measurement error is not the sole source, or even the most potent source, of unreliability affecting inferences drawn from class means. The test results for any given year reflect not only the character of the instructional program but also the character of students enrolled at that specific moment. These individuals must be regarded as a sample, in a longitudinal sense, from the population that flows through the district schools over a period of years. (p.127)

The decrease in measurement error from aggregation may be offset by sampling error that is not accounted for in traditional estimates of reliability. Brennan (1995) recommended using generalizability theory to account for sources of error from both items and samples of students. The application of generalizability theory to multilevel validity studies has implications for the interpretations of different sources of error at multiple levels of analysis.

The high between-level factor loadings support the use of student achievement tests for making school-level inferences in this study. The majority of items appeared more discriminating at the school level than at the student level. Contrary to the original concern that several items might be less appropriate for school-level inferences than for student-level inferences, it appears that the opposite scenario might apply. That is, if school-level inferences are of primary interest, some items that failed to discriminate at the student level but had adequate properties at the school level might have been unnecessarily removed from draft test forms at an early stage of the test development process. This finding provides an additional reason for incorporating a multilevel factor structure into studies of school level achievement; not only does this approach provide more appropriate validity evidence, but it could facilitate the test development process if criteria for item selection are less strict at the school level than at the student level of analysis.

There were a small number of items that had nonsignificant or negative factor loadings at the school level. Although the corresponding factor loadings at the student level were higher, the differences tended to be slight. There were no instances where an item was highly discriminating at the student level but had a very low factor loading at the school level. Instead, items with school-level factor loadings that were negative or close to zero had corresponding student-level factor loadings that were positive but very low. For example, Item 41 in grade 5 had a between-level factor loading of .10 and a within-level factor loading of .14. Item 7 in grade 8 had a between-level factor loading of -.29; the corresponding within-level factor loading was .05. From a practical standpoint, these scenarios do not lead to different conclusions about item quality despite the fact that both student-level factor loadings were positive and significant and the school-level factor loadings were not. None of

the factor loadings for these two items would provide strong evidence for including the items on a test; the only difference in conclusions of item quality here might be between "marginal" and "unacceptable."

The finding of one item with a significant negative factor loading (Item 7 in grade 8) suggests that the inclusion of this item detracts from the measurement of school-level achievement. The practical implications of including this item in a test used to draw school-level inferences are unclear, especially given the relatively low strength of the item (-.29). It is unlikely that rescoring the tests with the exclusion of this item would lead to different conclusions, but this is certainly an area for further research. Additional or larger negative factor loadings at the school level certainly could have practical implications if such items were included when drawing school-level inferences.

This particular finding of a significant negative factor loading at the school level may also yield valuable information about instruction or curricula. Given the relatively high item p-value for Item 7 in grade 8 (0.69), this finding suggests that students in low achieving schools may have been more likely to answer the item correctly than students in high achieving schools. Just as differential item functioning analyses can provide insight about students' strengths and weaknesses despite its primary use as identifying items that are potentially biased (Stanat & Lüdtke, 2008), multilevel confirmatory factor analyses have the potential to provide valuable information about school-level strengths and weaknesses that could be related to instruction or curricula. In future research, it would be interesting to examine the characteristics of items that show differential discrimination at the student and school levels.

Implications of Multilevel Validity Research

The specific research findings in this study are certainly noteworthy, but perhaps the most important implication of this research is its broader role in illustrating that multilevel validity analyses can and should be undertaken. This study can serve as a prototype not for the exact processes to be undertaken but for a type of general approach to gathering multilevel validity evidence. As both the methodological capabilities and research base in this field develop, it is likely that accepted standards for performing multilevel validity studies will evolve from the general approach taken and specific analyses performed here. The intent of this study was to serve as an initial attempt for sparking dialogue on research and practice in this area. The *Standards for Educational and Psychological Testing* (AERA et al., 1999) is clear that secondary uses of tests require that additional validity evidence be gathered to investigate the validity of those interpretations. What is less clear about using student achievement tests to draw school-level inferences is who is in the best position to collect such evidence. The primary purpose of state achievement test programs is to measure student achievement. Unless measuring school-level outcomes is an overt goal of state testing programs, they would not appear to be technically responsible for considering the school-level factor structure during test development or validation processes.

The *Standards* (AERA et al., 1999) places the impetus on the test user to collect additional validity evidence for secondary purposes. Although this guideline sounds reasonable in theory, it presents many practical challenges for collecting validity evidence related to school-level inferences from student achievement tests. First, access to item-level data (necessary to perform multilevel factor analyses) is generally restricted; only raw or scale scores tend to be publicly available and accessible from state departments of education.

Second, many educational researchers and program evaluators may lack the technical expertise necessary to perform such analyses, particularly now when the methodology is still in its infancy and clear guidelines for collecting multilevel validity evidence do not yet exist. If the impetus to collect multilevel validity evidence is placed solely upon educational researchers and program evaluators, much of the attractiveness of using state achievement data for secondary purposes would be diminished.

Given the current prominence of educational testing, it seems clear that state achievement tests will continue to be used to make school-level inferences, even if that is not the primary purpose of the tests. State testing programs may want to consider either taking preliminary steps to collect validity evidence for school-level inferences, or alternatively, explicitly stating that such claims have not been investigated and must be undertaken by secondary users before it is appropriate to use the tests for this purpose. According to Standard 1.3, "If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations" (AERA et al., 1999, p. 18). It is unlikely that state testing programs are currently collecting multilevel validity evidence or explicitly urging secondary users to do so, and it is even less likely that educational researchers and program evaluators are aware that such steps should be taken before the student tests are used for the secondary purpose of drawing school-level inferences.

Multilevel validity considerations would be most easily addressed during the test development process. One aspect of item analyses could be an examination of multilevel factor loadings. Based on the results in this study, few decisions about item quality are likely

to be affected, but if school-level inferences are intended, the small number of items with low or negative school-level loadings could be discarded. Of course, given the cost of developing even a single test item, such a step may not be feasible if state testing programs decide that school-level inferences are not an intended use of their tests, even as a secondary purpose. If the latter stance is taken, however, it should be accompanied by explicit warnings to secondary test users, as well as greater accessibility to the item-level data needed to perform additional analyses.

Regardless of the stance taken by state testing programs, the psychometric community should develop clearer guidance on multilevel validation. It is not enough for the *Standards* (AERA et al., 1999) to state that additional validity evidence is needed for secondary purposes, particularly when tests are used for the measurement of groups rather than individuals. As Linn (2006) suggested, there is a need for explicit guidance on psychometric issues specific to group-level measurement; this would certainly include multilevel validation efforts. Specific guidance on collecting multilevel validity evidence currently does not exist because it is deemed to fall outside the traditional boundaries of both psychometrics and program evaluation.

The discussion of multilevel validity thus far has focused on secondary uses of tests that were created originally for the primary purpose of making inferences about individuals. Although the responsibility for collecting such evidence may be muddled for secondary uses, it seems clear that test developers must consider multilevel validity when group-level inferences are a primary test purpose. For example, there are several educational tests administered to students for the primary purpose of making inferences at higher levels of aggregation, such as the National Assessment of Educational Progress (NAEP), Programme

for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMSS). Such tests are given to samples of students for the purpose of comparing schools, states, or countries.

The chapter in *Educational Measurement* (Brennan, 2006) on group-score assessments (Mazzeo et al., 2006) includes a discussion of how sampling, reliability, and scoring are affected by the test purpose, but consideration of multilevel validity is noticeably absent. Contrary to the claim that item analyses, "are not different from approaches used in individual-score tests, so are not discussed here" (Mazzeo et al., 2006), it seems that item analyses do warrant rethinking in the context of group-level inferences. In fact, a more recent chapter on group-level measurement in student achievement (Stanat & Lüdtke, 2008) included the following statement in regard to the PISA literacy tests:

> [T]he factor structure of the literacy tests at the individual student level and at the country level could be compared. To our knowledge, this analysis has not yet been performed, even though previous research has shown that a different factor structure may emerge already on the class or school level as compared to the individual level…. Due to methodological advances in the integration of structural equation and multilevel modeling, simultaneous analyses of factor structures and relationships among variables at different levels within a single model should become more prevalent in the future (p. 329).

For tests that are designed with the primary purpose of making group-level inferences, multilevel validity evidence may involve an additional layer of complexity. Often these group-score assessments include a matrix sampling design, where each student receives only a small sample of items and not a complete test form. Although this design is more efficient for drawing group-level inferences, current methods in multilevel factor analysis require that students respond to all items. Collecting multilevel validity evidence may require performing

a pilot or field test where samples of students do receive the entire test, although advances in methodology may eventually make such a requirement obsolete.

Other fields have more readily embraced multilevel validation efforts when group-level inferences are the primary test purpose, in particular the field of industrial and organizational psychology. For example, the book *Multi-level Issues in Organizational Behavior and Processes* (Yammarino & Dansereu, 2004) includes an entire section on construct validation, in which different approaches to reliability and validity in a multilevel framework are considered. The consideration of multilevel factor structure in this study is only one approach to investigating multilevel validity evidence. Future research in this area is needed to determine what aspects of multilevel validation are most relevant to educational achievement.

Increased attention to multilevel validity issues has the potential to spur more research and practice in innovative multilevel analyses. A new type of multilevel modeling that has not yet been embraced in education is known as a micro-macro situation (Croon & van Veldhoven, 2007). In traditional multilevel modeling (macro-micro situations), the dependent variable is measured at the lowest level and may be predicted by variables at that level or a higher level of aggregation. Conversely, in micro-macro situations, the dependent variable is actually measured at the highest level and may be predicted by variables at that level or at a lower level (Croon & van Veldhoven, 2007). In education, this would mean that a variable collected from a teacher or principal could be used as an outcome that is predicted or explained in part by student-level variables. As Croon and van Veldhoven (2007) point out, traditional software packages for performing HLM analyses (e.g., Raudenbush et al., 2004) do not allow for this type of modeling, but this novel approach offers a new way to think

about the measurement of group-level outcomes. The relationship between advances in multilevel validation and multilevel analyses is certainly an area worthy of future research.

Future Directions in Multilevel Psychometrics

Once the psychometric implications of multilevel modeling are considered, it becomes clear that item analyses and validity evidence may not be the only aspects of testing that are affected. One area of further research is to explore the implications of accounting for the multilevel data structure during the test scoring process. Multilevel item response theory models were originally developed in the context of matrix sampling, when each student was administered a single item and the goal was to produce estimates for a school (Mislevy, 1983). More recently, multilevel item response theory has been extended to instances where individual level abilities are also of interest (e.g., Kamata, 2001). Research in this area could consider the implications of accounting for the multilevel structure when producing student or school level scores.

Another emerging area in multilevel psychometrics is the application of multilevel differential item functioning (DIF). Cheong (2006) used a hierarchical generalized linear model to identify both school and student sources of differential item functioning in the U.S. Civic Education Study of the International Association for the Evaluation of Educational Achievement. Some items that were flagged for racial-ethnic bias on the student level no longer exhibited DIF when teacher-reported opportunity to learn was considered (Cheong, 2006). Multilevel DIF enables sources of variation on several different levels to be considered and is a ripe area for future research in multilevel psychometrics and student achievement.

Another potential application of multilevel psychometrics is multilevel standard setting. No examples of research or practice in this area could be found, but this appears to be an area worth considering in the context of multilevel test development. Multilevel applications of standard setting could take the form of using multilevel item response models in traditional standard setting processes, or even attempting to set standards at a higher level of aggregation that is of interest. The latter practice could be used to address some current concerns about accountability systems. In state achievement testing programs, standards are set in regard to student levels of performance; the standards are then aggregated to the school level for accountability purposes. It seems worthwhile to explore whether more meaningful school-level classifications could be developed directly using a multilevel framework.

Conclusion

Sirotnik (1980) called attention to the importance of multilevel psychometrics nearly 30 years ago, at a time when much of the technological advances in multilevel modeling had not yet occurred. Methodology in this area has developed to the point where many theoretical aspects of multilevel psychometrics can begin to be put into practice, but multilevel models have almost exclusively been applied at the analysis stage.

Although the current seminal references in psychometrics (i.e., AERA et al., 1999; Brennan, 2006) are largely silent on this issue, it appears that some psychometricians are beginning to take note of the importance of multilevel psychometrics. In an invited session at AERA/NCME on "The Big Challenges and Research Opportunities in Testing and Measurement" Zumbo (2008) gave a presentation entitled, "Testing and Measurement from a Multi-level View: Psychometrics and Validation." He called for consideration of the multilevel data structure during test development and validation when constructs are

interpreted and used at higher levels of aggregation. Zumbo and Forer (in press) elaborated on the importance of multilevel measurement procedures for tests that are designed exclusively to measure group-level constructs, such as NAEP. Linn (2008) recently noted the importance of considering multilevel validity evidence when using student achievement tests as measures of school quality in accountability systems. It seems plausible that future editions of the *Standards* (AERA et al., 1999) and *Educational Measurement* (Brennan, 2006) may begin to incorporate the notion of multilevel psychometrics as the concept gains favor in both theory and practice.

Current technological capabilities for performing multilevel psychometric analyses are rapidly evolving but are still very limited in comparison to traditional (single level) procedures. However, it is important to note that the research necessary for finding methodological solutions is largely driven by the awareness of existing problems. There is some evidence that the psychometric community is beginning to take note of the psychometric implications of multilevel data structures, and the seminal works in the field may soon be outdated in regard to multilevel test development. It is studies such as the one reported here that have the potential to generate both research and practice so that the good works done in the name of multilevel validity are no longer remarkably few.

APPENDIX A

Table A.1

Descriptive Statistics for Grade 3 Items

Item	Mean (SD)	Item	Mean (SD)	Item	Mean (SD)
1	0.58 (0.49)	21	0.78 (0.42)	41	0.55 (0.50)
2	0.95 (0.22)	22	0.71 (0.46)	42	0.77 (0.42)
3	0.48 (0.50)	23	0.55 (0.50)	43	0.87 (0.34)
4	0.61 (0.49)	24	0.80 (0.40)	44	0.86 (0.35)
5	0.47 (0.50)	25	0.87 (0.33)	45	0.85 (0.35)
6	0.79 (0.41)	26[1]	1.42 (0.85)	46	0.74 (0.44)
7	0.33 (0.47)	27[1]	1.06 (0.92)	47	0.86 (0.35)
8	0.48 (0.50)	28[1]	1.46 (0.81)	48	0.97 (0.18)
9	0.55 (0.50)	29[1]	1.23 (0.89)	49	0.72 (0.45)
10	0.84 (0.37)	30[1]	1.15 (0.91)	50	0.81 (0.40)
11	0.75 (0.43)	31[1]	1.13 (0.81)	51	0.78 (0.42)
12[1]	1.17 (0.90)	32[1]	1.49 (0.75)	52	0.60 (0.49)
13[1]	0.91 (0.90)	33[1]	1.19 (0.54)	53	0.91 (0.29)
14[1]	1.44 (0.70)	34[1]	1.51 (0.73)	54	0.36 (0.48)
15[1]	0.96 (0.83)	35	0.91 (0.29)	55	0.93 (0.25)
16[1]	0.72 (0.89)	36	0.94 (0.24)	56	0.67 (0.47)
17	0.65 (0.48)	37	0.89 (0.31)	57	0.78 (0.42)
18	0.76 (0.43)	38	0.81 (0.39)	58	0.85 (0.36)
19	0.51 (0.50)	39	0.68 (0.47)	59	0.71 (0.45)
20	0.68 (0.47)	40	0.79 (0.41)		

Note. SD = Standard Deviation.
[1] The range for these items was 0-2.

Table A.2

Descriptive Statistics for Grade 5 Items

Item	Mean (SD)	Item	Mean (SD)	Item	Mean (SD)
1	0.53 (0.50)	21	0.47 (0.50)	41	0.24 (0.43)
2	0.57 (0.50)	22	0.76 (0.43)	42	0.49 (0.50)
3	0.62 (0.49)	23	0.68 (0.47)	43	0.58 (0.49)
4	0.58 (0.49)	24	0.42 (0.49)	44	0.69 (0.46)
5	0.65 (0.48)	25	0.42 (0.49)	45	0.81 (0.39)
6	0.28 (0.45)	26[1]	1.04 (0.83)	46	0.59 (0.49)
7	0.39 (0.49)	27[1]	1.47 (0.79)	47	0.75 (0.43)
8	0.60 (0.49)	28[1]	0.51 (0.76)	48	0.80 (0.40)
9	0.58 (0.49)	29[1]	0.80 (0.86)	49	0.84 (0.37)
10	0.75 (0.43)	30[2]	1.81 (1.40)	50	0.95 (0.21)
11[1]	0.89 (0.88)	31[2]	1.77 (1.36)	51	0.50 (0.50)
12[1]	1.30 (0.77)	32	0.61 (0.49)	52	0.77 (0.42)
13[1]	1.39 (0.77)	33	0.62 (0.49)	53	0.49 (0.50)
14[1]	1.02 (0.78)	34	0.85 (0.36)	54	0.69 (0.46)
15[2]	2.11 (1.63)	35	0.66 (0.47)	55	0.85 (0.36)
16	0.79 (0.41)	36	0.62 (0.49)	56	0.76 (0.43)
17	0.62 (0.49)	37	0.75 (0.44)	57	0.49 (0.50)
18	0.78 (0.41)	38	0.87 (0.34)	58	0.59 (0.49)
19	0.72 (0.45)	39	0.81 (0.39)	59	0.47 (0.50)
20	0.75 (0.44)	40	0.70 (0.46)		

Note. SD = Standard Deviation.
[1] The range for these items was 0-2.
[2] The range for these items was 0-4.

Table A.3

Descriptive Statistics for Grade 8 Items

Item	Mean (SD)	Item	Mean (SD)	Item	Mean (SD)
1	0.77 (0.42)	21	0.78 (0.41)	41	0.35 (0.48)
2	0.45 (0.50)	22	0.45 (0.50)	42	0.59 (0.49)
3	0.43 (0.50)	23	0.65 (0.48)	43	0.84 (0.37)
4	0.65 (0.48)	24	0.37 (0.48)	44	0.62 (0.48)
5	0.48 (0.50)	25	0.79 (0.41)	45	0.63 (0.48)
6	0.48 (0.50)	26[1]	0.71 (0.90)	46	0.49 (0.50)
7	0.69 (0.46)	27[1]	0.56 (0.68)	47	0.48 (0.50)
8	0.20 (0.40)	28[1]	1.34 (0.82)	48	0.60 (0.49)
9	0.56 (0.50)	29[1]	1.13 (0.88)	49	0.81 (0.39)
10	0.27 (0.45)	30[2]	2.46 (1.38)	50	0.84 (0.36)
11[1]	0.88 (0.91)	31[2]	2.47 (1.52)	51	0.70 (0.46)
12[1]	0.77 (0.93)	32	0.21 (0.41)	52	0.52 (0.50)
13[1]	1.25 (0.75)	33	0.50 (0.50)	53	0.49 (0.50)
14[1]	0.95 (0.90)	34	0.21 (0.41)	54	0.68 (0.47)
15[2]	1.25 (1.27)	35	0.41 (0.49)	55	0.67 (0.47)
16	0.43 (0.50)	36	0.76 (0.43)	56	0.54 (0.50)
17	0.35 (0.48)	37	0.55 (0.50)	57	0.74 (0.44)
18	0.38 (0.49)	38	0.51 (0.50)	58	0.37 (0.48)
19	0.58 (0.49)	39	0.64 (0.48)	59	0.48 (0.50)
20	0.47 (0.50)	40	0.54 (0.50)		

Note. SD = Standard Deviation.
[1] The range for these items was 0-2.
[2] The range for these items was 0-4.

APPENDIX B

Table B.1.

Two-level Exploratory Factor Analyses Fit Statistics for Grade 3, All Models

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	2,855.09* (1652)	1.00	1.00	.01	.00	.06
UN	2	2,169.19* (1594)	1.00	1.00	.01	.00	.05
UN	3	1,665.30* (1537)	1.00	1.00	.00	.00	.04
1	UN	7,876.13* (1652)	.99	.98	.02	.03	.00
2	UN	5,495.40* (1594)	.99	.99	.02	.03	.00
3	UN	3,798.30* (1537)	1.00	.99	.01	.02	.00
1	1	14,374.00* (3304)	.98	.98	.02	.03	.06
1	2	14,153.34* (3246)	.98	.98	.02	.03	.05
1	3	13,983.62* (3189)	.98	.98	.02	.03	.04
2	1	10,377.74* (3246)	.99	.99	.02	.03	.06
2	2	10,145.22* (3188)	.99	.99	.02	.03	.05
2	3	9,976.14* (3131)	.99	.99	.02	.03	.04
3	1	7,517.76* (3189)	.99	.99	.01	.02	.06
3	2	7,272.16* (3131)	.99	.99	.01	.02	.05
3	3	7,099.61* (3074)	.99	.99	.01	.02	.04

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

Table B.2.

Two-level Exploratory Factor Analyses Fit Statistics for Grade 5, All Models

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	2,255.57* (1652)	1.00	1.00	.01	.00	.10
UN	2	1,928.12* (1594)	1.00	1.00	.01	.00	.09
UN	3	1,641.61* (1537)	1.00	1.00	.00	.00	.08
1	UN	5,551.48* (1652)	.99	.98	.02	.03	.00
2	UN	4,059.28* (1594)	.99	.99	.01	.02	.00
3	UN	3,255.25* (1537)	1.00	.99	.01	.02	.00
1	1	9,649.71* (3304)	.99	.99	.02	.03	.10
1	2	9,557.64* (3246)	.99	.99	.02	.03	.09
1	3	9.488.26* (3189)	.99	.99	.02	.03	.08
2	1	7,385.96* (3246)	.99	.99	.01	.02	.10
2	2	7,258.78* (3188)	.99	.99	.01	.02	.09
2	3	7,161.18* (3131)	.99	.99	.01	.02	.08
3	1	6181.12* (3189)	.99	.99	.01	.02	.10
3	2	6033.84* (3131)	.99	.99	.01	.02	.09
3	3	5,920.06* (3074)	.99	.99	.01	.02	.08

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

Table B.3.

Two-level Exploratory Factor Analyses Fit Statistics for Grade 8, All Models

Within Level Factors	Between Level Factors	χ^2 (df)	CFI	TLI	RMSEA	SRMR (within)	SRMR (between)
UN	1	1,068.96 (1652)	1.00	1.00	.00	.00	.05
UN	2	573.51 (1594)	1.00	1.00	.00	.00	.04
UN	3	393.35 (1537)	1.00	1.01	.00	.00	.03
1	UN	9,204.67* (1652)	.98	.97	.02	.04	.00
2	UN	2,969.45* (1594)	1.00	.99	.01	.02	.00
3	UN	2,205.78* (1537)	1.00	1.00	.01	.02	.00
1	1	16,200.54* (3304)	.97	.97	.02	.04	.05
1	2	16,194.27* (3246)	.97	.97	.02	.04	.04
1	3	16,079.41* (3189)	.97	.97	.02	.04	.03
2	1	5,481.39* (3246)	1.00	1.00	.01	.02	.05
2	2	5,372.34* (3188)	1.00	1.00	.01	.02	.04
2	3	5,302.24* (3131)	1.00	1.00	.01	.02	.03
3	1	4,204.20* (3189)	1.00	1.00	.01	.02	.05
3	2	4,081.33* (3131)	1.00	1.00	.01	.02	.04
3	3	4,015.77* (3074)	1.00	1.00	.01	.02	.03

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Squared Residual; UN = unrestricted.
* $p < .05$

APPENDIX C

Table C.1

Two-level Exploratory Factor Analysis Solution (Two Factors at Each Level) for Grade 3

	Within-level Loadings (SE)		Between-level Loadings (SE)	
Item	Factor 1	Factor 2	Factor 1	Factor 2
1	.58* (.02)	.14* (.03)	.86* (.04)	.24* (.08)
2	.26* (.02)	.02 (.04)	.80* (.09)	.26* (.13)
3	.33* (.03)	.27 (.02)	.79* (.08)	.32* (.10)
4	.65* (.02)	.11* (.03)	.95* (.02)	.09 (.07)
5	.60* (.03)	.25* (.03)	.85* (.05)	.30* (.08)
6	.66* (.01)	-.02 (.03)	.95* (.02)	-.05 (.06)
7	.45* (.03)	.31* (.03)	.82* (.06)	.37* (.07)
8	.60* (.03)	.20* (.03)	.91* (.03)	.12* (.06)
9	.49* (.02)	.17* (.02)	.95* (.03)	.07* (.06)
10	.57* (.02)	-.09* (.03)	.80* (.07)	-.42* (.08)
11	.45* (.01)	.05 (.03)	.86* (.04)	-.10 (.07)
12	.63* (.01)	.07* (.02)	.94* (.03)	.19* (.06)
13	.70* (.02)	.12* (.03)	.93* (.03)	.18* (.06)
14	.60* (.01)	-.15* (.03)	.99* (.01)	-.04 (.06)
15	.46* (.02)	.10* (.03)	.92* (.04)	.27* (.07)
16	.70* (.02)	.21* (.03)	.96* (.03)	.21* (.05)
17	.57* (.02)	.16* (.03)	.84* (.04)	.08 (.08)
18	.45* (.01)	.05 (.03)	.95* (.03)	-.13 (.08)
19	.61* (.03)	.20* (.03)	.97* (.03)	.17* (.05)
20	.63* (.01)	.03 (.03)	.95* (.02)	.02 (.06)

Note. SE = Standard Error.
* $p < .05$

Table C.1 (continued)

Two-level Exploratory Factor Analysis Solution (Two Factors at Each Level) for Grade 3

	Within-level Loadings (SE)		Between-level Loadings (SE)	
Item	Factor 1	Factor 2	Factor 1	Factor 2
21	.54* (.01)	-.04 (.03)	.87 (.03)	-.05 (.07)
22	.26* (.01)	.01 (.03)	.87 (.05)	-.05 (.08)
23	.53* (.02)	.10* (.02)	.84 (.04)	.22 (.08)
24	.53* (.01)	-.06* (.03)	.93 (.02)	-.04 (.07)
25	.46* (.02)	.01 (.03)	.90 (.03)	-.06 (.07)
26	.64* (.01)	.00 (.01)	.92 (.02)	.04 (.05)
27	.65* (.01)	.10* (.02)	.92 (.04)	.26 (.06)
28	.59* (.01)	.02 (.03)	.91 (.03)	.15 (.07)
29	.63* (.01)	.09 (.02)	.93 (.03)	.17 (.06)
30	.44* (.02)	.10* (.02)	.79 (.04)	-.06 (.07)
31	.55* (.01)	.01 (.02)	.92 (.02)	-.01 (.04)
32	.58* (.01)	-.01 (.02)	.97 (.02)	.07 (.05)
33	.39* (.01)	-.03 (.02)	.95 (.03)	.11 (.08)
34	.54* (.01)	-.07* (.02)	.91 (.03)	.01 (.04)
35	.74* (.03)	-.23* (.04)	.90 (.04)	-.26 (.07)
36	.32* (.02)	-.11* (.04)	.73 (.09)	.18 (.11)
37	.46* (.03)	-.17* (.03)	.83 (.07)	-.12 (.10)
38	.65* (.02)	-.08* (.03)	.98 (.03)	-.11 (.06)
39	.55* (.01)	-.03 (.03)	.79 (.05)	-.07 (.07)
40	.56* (.01)	-.02 (.03)	.91 (.03)	.07 (.07)

Note. SE = Standard Error.
* $p < .05$

Table C.1 (continued)

Two-level Exploratory Factor Analysis Solution (Two Factors at Each Level) for Grade 3

	Within-level Loadings (SE)		Between-level Loadings (SE)	
Item	Factor 1	Factor 2	Factor 1	Factor 2
41	.43* (.02)	.07* (.02)	.91* (.04)	.05 (.09)
42	.53* (.02)	-.10* (.03)	.92* (.04)	-.19* (.07)
43	.62* (.03)	-.16* (.03)	.75* (.05)	-.10 (.08)
44	.59* (.03)	-.25* (.03)	.71* (.08)	-.40* (.07)
45	.62* (.02)	-.12* (.03)	.97* (.02)	-.09 (.07)
46	.69* (.01)	.03 (.03)	.94* (.02)	.01 (.05)
47	.66* (.04)	-.33* (.03)	.89* (.05)	-.24* (.07)
48	.50* (.05)	-.33* (.05)	.96* (.07)	-.08 (.09)
49	.62* (.01)	-.05* (.02)	.98* (.02)	-.02 (.06)
50	.56* (.01)	-.05 (.03)	.95* (.03)	-.10 (.07)
51	.50* (.02)	-.13* (.02)	.88* (.06)	-.18* (.09)
52	.47* (.02)	.18* (.03)	.91* (.04)	.10 (.08)
53	.34* (.02)	-.08* (.03)	.93* (.06)	-.27* (.09)
54	.40* (.02)	.15* (.03)	.81* (.04)	.09 (.09)
55	.47* (.03)	-.22* (.04)	.63* (.10)	-.53* (.10)
56	.53* (.01)	.02 (.02)	.91* (.03)	-.04 (.07)
57	.48* (.02)	-.13* (.03)	.86* (.05)	-.23* (.07)
58	.66* (.01)	-.07* (.03)	.97* (.02)	-.01 (.05)
59	.50* (.02)	-.10* (.03)	.96* (.05)	-.27* (.10)

Note. SE = Standard Error.
* $p < .05$

APPENDIX D

Table D.1

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
1	.58* (.01)	.85* (.04)
2	.26* (.02)	.79* (.08)
3	.34* (.01)	.78* (.06)
4	.65* (.01)	.94* (.02)
5	.61* (.01)	.84* (.04)
6	.66* (.01)	.95* (.02)
7	.47* (.01)	.80* (.04)
8	.61* (.01)	.91* (.03)
9	.50* (.01)	.95* (.03)
10	.57* (.01)	.80* (.04)
11	.46* (.01)	.87* (.03)
12	.63* (.01)	.93* (.02)
13	.71* (.01)	.92* (.02)
14	.58* (.01)	.99* (.01)
15	.46* (.01)	.91* (.03)
16	.72* (.01)	.95* (.02)
17	.58* (.01)	.84* (.04)
18	.46* (.01)	.96* (.03)
19	.62* (.01)	.96* (.02)
20	.63* (.01)	.95* (.02)

Note. SE = Standard Error.
* $p < .05$

Table D.1 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
21	.54* (.01)	.87* (.03)
22	.26* (.01)	.87* (.04)
23	.54* (.01)	.83* (.04)
24	.53* (.01)	.94* (.02)
25	.46* (.02)	.91* (.03)
26	.64* (.01)	.92* (.02)
27	.65* (.01)	.91* (.02)
28	.59* (.01)	.90* (.03)
29	.64* (.01)	.92* (.02)
30	.44* (.01)	.80* (.04)
31	.54* (.01)	.92* (.02)
32	.59* (.01)	.96* (.01)
33	.39* (.01)	.94* (.03)
34	.53* (.01)	.91* (.03)
35	.73* (.01)	.91* (.03)
36	.32* (.02)	.73* (.09)
37	.46* (.02)	.83* (.06)
38	.65* (.01)	.98* (.02)
39	.55* (.01)	.80* (.04)
40	.56* (.01)	.91* (.03)

Note. SE = Standard Error.
* $p < .05$

Table D.1 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
41	.43* (.01)	.91* (.04)
42	.52* (.01)	.92* (.03)
43	.61* (.01)	.76* (.05)
44	.58* (.01)	.72* (.06)
45	.61* (.01)	.98* (.02)
46	.69* (.01)	.94* (.02)
47	.65* (.01)	.90* (.03)
48	.50* (.02)	.96* (.07)
49	.61* (.01)	.98* (.02)
50	.56* (.01)	.96* (.02)
51	.50* (.01)	.88* (.04)
52	.48* (.01)	.91* (.03)
53	.34* (.02)	.93* (.05)
54	.41* (.01)	.81* (.04)
55	.46* (.02)	.63* (.07)
56	.53* (.01)	.91* (.03)
57	.48* (.01)	.87* (.03)
58	.65* (.01)	.97* (.02)
59	.50* (.01)	.96* (.05)

Note. SE = Standard Error.
* $p < .05$

Table D.2

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 5

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
1	.55* (.01)	.96* (.03)
2	.49* (.01)	.98* (.03)
3	.47* (.01)	.96* (.03)
4	.41* (.01)	.94* (.04)
5	.48* (.01)	.81* (.05)
6	-.02 (.02)	-.13 (.20)
7	.62* (.01)	.93* (.03)
8	.56* (.01)	.99* (.02)
9	.36* (.01)	.86* (.07)
10	.45* (.01)	.77* (.05)
11	.56* (.01)	.94* (.02)
12	.36* (.01)	.86* (.05)
13	.62* (.01)	.93* (.02)
14	.54* (.01)	.71* (.06)
15	.61* (.01)	.91* (.03)
16	.30* (.01)	.98* (.03)
17	.63* (.01)	.97* (.02)
18	.44* (.01)	.54* (.09)
19	.58* (.01)	.90* (.03)
20	.61* (.01)	.90* (.03)

Note. SE = Standard Error.
* $p < .05$

Table D.2 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 5

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
21	.59* (.01)	.88* (.03)
22	.56* (.01)	.91* (.04)
23	.51* (.01)	.93* (.03)
24	.39* (.01)	.56* (.07)
25	.46* (.01)	.96* (.03)
26	.50* (.01)	.80* (.05)
27	.39* (.01)	.76* (.07)
28	.62* (.01)	.72* (.05)
29	.55* (.01)	.94* (.02)
30	.55* (.01)	.97* (.01)
31	.59* (.01)	.93* (.03)
32	.37* (.01)	.71* (.08)
33	.46* (.01)	.24* (.11)
34	.52* (.01)	.94* (.03)
35	.50* (.01)	.59* (.08)
36	.63* (.01)	.91* (.03)
37	.51* (.01)	.89* (.04)
38	.43* (.02)	.82* (.06)
39	.28* (.02)	.77* (.11)
40	.41* (.01)	.77* (.06)

Note. SE = Standard Error.
* $p < .05$

Table D.2 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 5

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
41	.14* (.02)	.10 (.17)
42	.43* (.01)	.53* (.08)
43	.59* (.01)	.83* (.05)
44	.56* (.01)	.91* (.03)
45	.64* (.01)	.98* (.03)
46	.57* (.01)	.94* (.03)
47	.49* (.01)	.82* (.06)
48	.39* (.02)	.76* (.06)
49	.38* (.02)	.88* (.06)
50	.52* (.02)	.94* (.09)
51	.47* (.01)	.84* (.04)
52	.62* (.01)	.96* (.02)
53	.49* (.01)	.72* (.06)
54	.53* (.01)	.82* (.04)
55	.40* (.02)	.95* (.06)
56	.28* (.02)	.73* (.08)
57	.56* (.01)	.96* (.03)
58	.52* (.01)	.95* (.03)
59	.15* (.01)	.68* (.19)

Note. SE = Standard Error.
* $p < .05$

Table D.3

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 8

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
1	.51* (.02)	.94* (.02)
2	.49* (.01)	.93* (.04)
3	.21* (.01)	.47* (.10)
4	.56* (.01)	.92* (.03)
5	.43* (.02)	.88* (.03)
6	.40* (.01)	.83* (.06)
7	.05* (.01)	-.29* (.12)
8	.34* (.01)	.69* (.08)
9	.51* (.01)	.88* (.04)
10	.57* (.02)	.87* (.05)
11	.67* (.01)	.96* (.05)
12	.70* (.02)	.95* (.08)
13	.56* (.01)	.97* (.02)
14	.73* (.01)	.96* (.06)
15	.64* (.01)	.93* (.06)
16	.45* (.01)	.97* (.02)
17	.47* (.01)	.97* (.02)
18	.55* (.01)	.99* (.01)
19	.50* (.01)	.98* (.01)
20	.43* (.01)	.88* (.04)

Note. SE = Standard Error.
* $p < .05$

Table D.3 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 8

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
21	.56* (.01)	.96* (.02)
22	.72* (.01)	.96* (.05)
23	.44* (.01)	.96* (.02)
24	.50* (.01)	.98* (.01)
25	.61* (.01)	.98* (.01)
26	.56* (.01)	.93* (.03)
27	.50* (.01)	.95* (.03)
28	.61* (.01)	.98* (.03)
29	.72* (.01)	.96* (.02)
30	.61* (.01)	.94* (.02)
31	.72* (.01)	.95* (.05)
32	.54* (.01)	.90* (.03)
33	.56* (.01)	1.00* (.01)
34	.67* (.02)	.71* (.08)
35	.68* (.01)	.98* (.02)
36	.52* (.01)	.89* (.03)
37	.48* (.01)	.96* (.02)
38	.55* (.01)	.96* (.02)
39	.23* (.02)	.68* (.07)
40	.64* (.01)	.97* (.02)

Note. SE = Standard Error.
* $p < .05$

Table D.3 (continued)

Two-level Exploratory Factor Analysis Solution (One Factor at Each Level) for Grade 8

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
41	.19* (.02)	.56* (.09)
42	.71* (.01)	.99* (.02)
43	.55* (.01)	.94* (.02)
44	.51* (.01)	.98* (.01)
45	.61* (.01)	.93* (.02)
46	.39* (.01)	.78* (.07)
47	.61* (.01)	.94* (.04)
48	.52* (.01)	.96* (.01)
49	.49* (.01)	.96* (.02)
50	.46* (.01)	.89* (.03)
51	.61* (.01)	.94* (.02)
52	.63* (.01)	.91* (.03)
53	.57* (.01)	.95* (.02)
54	.54* (.01)	.99* (.02)
55	.45* (.01)	.97* (.02)
56	.52* (.01)	.97* (.01)
57	.48* (.01)	.99* (.01)
58	.38* (.01)	.91* (.05)
59	.50* (.01)	.97* (.01)

Note. SE = Standard Error.
* $p < .05$

Appendix E

Table E.1

Multilevel Confirmatory Factor Analysis (Unstandardized Solution) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
1	1.00 (.00)	1.00 (.00)
2	.38* (.03)	.64* (.11)
3	.51* (.03)	.56* (.07)
4	1.20* (.04)	1.43* (.16)
5	1.07* (.03)	.99* (.11)
6	1.21* (.05)	1.22* (.02)
7	.75* (.03)	.83* (.10)
8	1.08* (.04)	1.29* (.03)
9	.80* (.03)	.90* (.08)
10	.97* (.04)	1.28* (.20)
11	.72* (.03)	1.05* (.16)
12	1.13* (.04)	1.43* (.17)
13	1.39* (.05)	2.04* (.26)
14	1.03* (.03)	1.28* (.16)
15	.73* (.03)	.96* (.13)
16	1.43* (.05)	1.82* (.22)
17	1.00* (.04)	1.16* (.15)
18	.72* (.03)	.74* (.11)
19	1.11* (.04)	1.10* (.13)
20	1.13* (.04)	1.05* (.11)

Note. SE = Standard Error.
* $p < .05$

Table E.1 (continued)

Multilevel Confirmatory Factor Analysis (Unstandardized Solution) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
21	.89* (.04)	1.10* (.14)
22	.38* (.02)	.65* (.11)
23	.89* (.03)	.91* (.12)
24	.87* (.04)	1.13* (.15)
25	.73* (.03)	1.37* (.19)
26	1.16* (.04)	1.87* (.22)
27	1.20* (.04)	1.40* (.18)
28	1.02* (.04)	1.29* (.17)
29	1.15* (.04)	1.38* (.19)
30	.69* (.03)	1.08* (.13)
31	.91* (.03)	1.41* (.16)
32	1.00* (.03)	1.51* (.20)
33	.59* (.02)	.77* (.09)
34	.89* (.03)	1.22* (.19)
35	1.52* (.07)	1.57* (.21)
36	.47* (.04)	.56* (.10)
37	.72* (.04)	.64* (.09)
38	1.18* (.05)	1.30* (.17)
39	.93* (.03)	1.32* (.16)
40	.94* (.04)	1.20* (.14)

Note. SE = Standard Error.
* $p < .05$

Table E.1 (continued)

Multilevel Confirmatory Factor Analysis (Unstandardized Solution) for Grade 3

Item	Within-level Loadings (SE)	Between-level Loadings (SE)
41	.67* (.03)	.63* (.09)
42	.86* (.04)	1.00* (.12)
43	1.09* (.05)	1.27* (.20)
44	1.00* (.04)	1.02* (.17)
45	1.08* (.05)	1.24* (.17)
46	1.32* (.04)	1.30* (.14)
47	1.19* (.05)	1.31* (.17)
48	.80* (.06)	.86* (.16)
49	1.09* (.04)	1.12* (.14)
50	.95* (.04)	1.20* (.13)
51	.81* (.04)	.97* (.15)
52	.76* (.03)	1.06* (.14)
53	.50* (.04)	.80* (.13)
54	.63* (.03)	.88* (.11)
55	.73* (.05)	1.04* (.19)
56	.87* (.03)	1.25* (.17)
57	.76* (.03)	1.02* (.13)
58	1.21* (.04)	1.38* (.18)
59	.80* (.03)	.59* (.08)

Note. SE = Standard Error.
$p < .05$

APPENDIX F

Table F.1

Confirmatory Factor Analyses Solutions for Grade 3 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.58* (.01)	.85* (.04)	.60* (.01)
2	.27* (.02)	.79* (.08)	.31* (.02)
3	.34* (.01)	.78* (.06)	.37* (.01)
4	.65* (.01)	.94* (.02)	.69* (.01)
5	.61* (.01)	.83* (.04)	.62* (.01)
6	.66* (.01)	.95* (.02)	.68* (.01)
7	.47* (.01)	.80* (.04)	.50* (.01)
8	.61* (.01)	.91* (.03)	.64* (.01)
9	.50* (.01)	.95* (.03)	.53* (.01)
10	.57* (.01)	.80* (.04)	.60* (.01)
11	.46* (.01)	.87* (.03)	.50* (.01)
12	.63* (.01)	.93* (.02)	.67* (.01)
13	.71* (.01)	.92* (.02)	.74* (.01)
14	.59* (.01)	.99* (.01)	.63* (.01)
15	.46* (.01)	.91* (.03)	.50* (.01)
16	.72* (.01)	.95* (.02)	.75* (.01)
17	.58* (.01)	.84* (.04)	.61* (.01)
18	.45* (.01)	.96* (.03)	.48* (.01)
19	.62* (.01)	.96* (.02)	.64* (.01)
20	.63* (.01)	.95* (.02)	.65* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.1 (continued)

Confirmatory Factor Analyses Solutions for Grade 3 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.54* (.01)	.87* (.03)	.57* (.01)
22	.26* (.01)	.87* (.04)	.31* (.01)
23	.54* (.01)	.83* (.04)	.56* (.01)
24	.53* (.01)	.94* (.02)	.57* (.01)
25	.46* (.02)	.91* (.03)	.53* (.01)
26	.64* (.01)	.92* (.02)	.69* (.01)
27	.65* (.01)	.91* (.02)	.68* (.01)
28	.59* (.01)	.90* (.03)	.63* (.01)
29	.64* (.01)	.92* (.02)	.67* (.01)
30	.44* (.01)	.80* (.04)	.49* (.01)
31	.54* (.01)	.92* (.02)	.60* (.01)
32	.58* (.01)	.96* (.01)	.64* (.01)
33	.39* (.01)	.95* (.03)	.43* (.01)
34	.54* (.01)	.91* (.03)	.58* (.01)
35	.73* (.01)	.90* (.03)	.75* (.01)
36	.32* (.02)	.73* (.09)	.35* (.02)
37	.46* (.02)	.83* (.06)	.48* (.02)
38	.65* (.01)	.98* (.02)	.68* (.01)
39	.55* (.01)	.80* (.04)	.59* (.01)
40	.56* (.01)	.91* (.03)	.60* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.1 (continued)

Confirmatory Factor Analyses Solutions for Grade 3 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.43* (.01)	.91* (.04)	.45* (.01)
42	.52* (.01)	.92* (.03)	.55* (.01)
43	.61* (.01)	.76* (.05)	.63* (.01)
44	.58* (.01)	.72* (.06)	.60* (.01)
45	.61* (.01)	.98* (.02)	.65* (.01)
46	.69* (.01)	.94* (.02)	.71* (.01)
47	.65* (.01)	.90* (.03)	.67* (.01)
48	.50* (.02)	.96* (.07)	.53* (.02)
49	.61* (.01)	.98* (.02)	.64* (.01)
50	.56* (.01)	.96* (.02)	.60* (.01)
51	.50* (.01)	.88* (.04)	.54* (.01)
52	.48* (.01)	.91* (.03)	.53* (.01)
53	.34* (.02)	.93* (.05)	.39* (.02)
54	.41* (.01)	.81* (.04)	.45* (.01)
55	.46* (.02)	.63* (.07)	.50* (.02)
56	.53* (.01)	.91* (.03)	.57* (.01)
57	.48* (.01)	.87* (.03)	.52* (.01)
58	.65* (.01)	.97* (.02)	.69* (.01)
59	.50* (.01)	.96* (.05)	.50* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.2

Confirmatory Factor Analyses Solutions for Grade 5 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.55* (.01)	.95* (.04)	.57* (.01)
2	.49* (.01)	.98* (.02)	.52* (.01)
3	.47* (.01)	.96* (.03)	.49* (.01)
4	.41* (.01)	.95* (.04)	.44* (.01)
5	.48* (.01)	.81* (.05)	.51* (.01)
6	-.02 (.02)	-.11 (.20)	-.02 (.02)
7	.62* (.01)	.93* (.03)	.64* (.01)
8	.56* (.01)	.99* (.02)	.59* (.01)
9	.36* (.01)	.86* (.07)	.38* (.01)
10	.45* (.01)	.77* (.05)	.47* (.01)
11	.56* (.01)	.94* (.02)	.58* (.01)
12	.36* (.01)	.86* (.06)	.38* (.01)
13	.62* (.01)	.93* (.02)	.63* (.01)
14	.54* (.01)	.70* (.06)	.56* (.01)
15	.60* (.01)	.90* (.03)	.63* (.01)
16	.30* (.01)	.98* (.03)	.33* (.01)
17	.62* (.01)	.97* (.02)	.65* (.01)
18	.45* (.01)	.55* (.09)	.46* (.01)
19	.58* (.01)	.90* (.03)	.62* (.01)
20	.61* (.01)	.89* (.04)	.63* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.2 (continued)

Confirmatory Factor Analyses Solutions for Grade 5 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.59* (.01)	.88* (.03)	.61* (.01)
22	.56* (.01)	.91* (.04)	.58* (.01)
23	.51* (.01)	.93* (.04)	.53* (.01)
24	.39* (.01)	.55* (.07)	.41* (.01)
25	.46* (.01)	.96* (.04)	.47* (.01)
26	.50* (.01)	.80* (.05)	.51* (.01)
27	.38* (.01)	.75* (.07)	.40* (.01)
28	.62* (.01)	.72* (.05)	.63* (.01)
29	.55* (.01)	.94* (.02)	.58* (.01)
30	.54* (.01)	.97* (.02)	.58* (.01)
31	.59* (.01)	.93* (.03)	.62* (.01)
32	.37* (.01)	.70* (.08)	.38* (.01)
33	.46* (.01)	.24* (.11)	.43* (.01)
34	.53* (.01)	.94* (.03)	.56* (.01)
35	.50* (.01)	.58* (.08)	.50* (.01)
36	.63* (.01)	.91* (.03)	.66* (.01)
37	.51* (.01)	.89* (.04)	.54* (.01)
38	.44* (.02)	.82* (.06)	.46* (.01)
39	.28* (.02)	.76* (.11)	.29* (.02)
40	.41* (.01)	.76* (.06)	.43* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.2 (continued)

Confirmatory Factor Analyses Solutions for Grade 5 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.14* (.02)	.12 (.17)	.14* (.02)
42	.43* (.01)	.53* (.08)	.44* (.01)
43	.59* (.01)	.84* (.05)	.60* (.01)
44	.56* (.01)	.90* (.04)	.59* (.01)
45	.64* (.01)	.98* (.03)	.65* (.01)
46	.57* (.01)	.94* (.03)	.59* (.01)
47	.49* (.01)	.82* (.06)	.50* (.01)
48	.39* (.02)	.75* (.06)	.41* (.01)
49	.38* (.02)	.88* (.06)	.40* (.02)
50	.51* (.02)	.94* (.09)	.53* (.02)
51	.48* (.01)	.84* (.04)	.51* (.01)
52	.62* (.01)	.96* (.02)	.65* (.01)
53	.49* (.01)	.72* (.06)	.51* (.01)
54	.53* (.01)	.82* (.04)	.56* (.01)
55	.39* (.02)	.95* (.06)	.42* (.02)
56	.28* (.02)	.72* (.08)	.31* (.01)
57	.56* (.01)	.96* (.03)	.59* (.01)
58	.52* (.01)	.95* (.03)	.55* (.01)
59	.15* (.01)	.68* (.19)	.16* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.3

Confirmatory Factor Analyses Solutions for Grade 8 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
1	.52* (.02)	.94* (.02)	.55* (.01)
2	.49* (.01)	.93* (.04)	.53* (.01)
3	.21* (.01)	.47* (.10)	.20* (.01)
4	.56* (.01)	.92* (.03)	.60* (.01)
5	.43* (.02)	.88* (.03)	.45* (.01)
6	.39* (.01)	.83* (.06)	.47* (.01)
7	.05* (.01)	-.31* (.12)	.02 (.01)
8	.35* (.01)	.70* (.08)	.39* (.02)
9	.50* (.01)	.88* (.04)	.56* (.01)
10	.57* (.02)	.87* (.05)	.58* (.01)
11	.67* (.01)	.96* (.05)	.73* (.01)
12	.69* (.02)	.95* (.08)	.72* (.01)
13	.56* (.01)	.97* (.02)	.60* (.01)
14	.73* (.01)	.96* (.06)	.76* (.01)
15	.64* (.01)	.93* (.06)	.68* (.01)
16	.45* (.01)	.97* (.02)	.48* (.01)
17	.47* (.01)	.97* (.02)	.50* (.01)
18	.55* (.01)	.99* (.01)	.59* (.01)
19	.49* (.01)	.98* (.01)	.53* (.01)
20	.42* (.01)	.87* (.04)	.44* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.3 (continued)

Confirmatory Factor Analyses Solutions for Grade 8 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
21	.55* (.01)	.96* (.02)	.60* (.01)
22	.72* (.01)	.97* (.05)	.76* (.01)
23	.44* (.01)	.96* (.02)	.47* (.01)
24	.50* (.01)	.98* (.01)	.55* (.01)
25	.61* (.01)	.98* (.01)	.66* (.01)
26	.55* (.01)	.93* (.03)	.61* (.01)
27	.50* (.01)	.95* (.03)	.55* (.01)
28	.61* (.01)	.98* (.03)	.68* (.01)
29	.72* (.01)	.96* (.02)	.76* (.01)
30	.62* (.01)	.95* (.02)	.66* (.01)
31	.73* (.01)	.95* (.05)	.74* (.01)
32	.54* (.01)	.91* (.03)	.56* (.01)
33	.55* (.01)	1.00* (.01)	.59* (.01)
34	.69* (.01)	.71* (.08)	.66* (.01)
35	.68* (.01)	.98* (.02)	.72* (.01)
36	.52* (.01)	.90* (.03)	.56* (.01)
37	.48* (.01)	.96* (.02)	.53* (.01)
38	.55* (.01)	.96* (.02)	.59* (.01)
39	.24* (.02)	.69* (.07)	.26* (.01)
40	.64* (.01)	.96* (.02)	.67* (.01)

Note. SE = Standard Error.
* $p < .05$

Table F.3 (continued)

Confirmatory Factor Analyses Solutions for Grade 8 (Limited to Clusters of Five or More)

	Multilevel CFA		Standard CFA
Item	Within-level Loadings (SE)	Between-level Loadings (SE)	Collapsed Loadings (SE)
41	.20* (.02)	.55* (.09)	.22* (.01)
42	.71* (.01)	.99* (.02)	.75* (.01)
43	.56* (.01)	.94* (.02)	.61* (.01)
44	.51* (.01)	.98* (.01)	.55* (.01)
45	.61* (.01)	.93* (.02)	.65* (.01)
46	.39* (.01)	.79* (.07)	.39* (.01)
47	.61* (.01)	.95* (.04)	.63* (.01)
48	.52* (.01)	.96* (.01)	.56* (.01)
49	.50* (.01)	.96* (.02)	.54* (.01)
50	.47* (.01)	.89* (.03)	.50* (.01)
51	.61* (.01)	.94* (.02)	.65* (.01)
52	.63* (.01)	.91* (.03)	.66* (.01)
53	.57* (.01)	.95* (.02)	.61* (.01)
54	.53* (.01)	.99* (.02)	.57* (.01)
55	.45* (.01)	.96* (.02)	.49* (.01)
56	.52* (.01)	.97* (.01)	.55* (.01)
57	.49* (.01)	.99* (.01)	.53* (.01)
58	.39* (.01)	.90* (.06)	.41* (.01)
59	.50* (.01)	.97* (.01)	.56* (.01)

Note. SE = Standard Error.
* $p < .05$

REFERENCES

Allodi, M. W. (2002). A two-level analysis of classroom climate in relation to social context, group composition, and organization of special support. *Learning Environments Research, 5,* 253-274.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

American Psychological Association (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: Author.

Asparouhov, T., & Muthén, B. (2007, August). Computationally efficient estimation of multilevel high-dimensional latent variable models. Paper presented at the annual meeting of the Joint Statistical Association (ASA section on Biometrics), Salt Lake City, UT.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Branum-Martin, L., Mehta, P. D., Carlson, C. D., Carlo, M., Fletcher, J. M., Ortiz, A., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*(1), 170-181.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*(4), 385-396.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5-9, 30.

Brennan, R. L (2006). (Ed.). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the *Standards for Educational and Psychological Testing. Educational Measurement: Issues and Practice, 25*(3), 35-41.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Cerin, E., Leslie, E., Owen, N., & Bauman, A. (2008). An Australian version of the Neighborhood environment walkability scale: Validity evidence. *Measurement in Physical Education, 12,* 31-51.

Cerin, E., Saelens, B. E., Sallis, J. F., & Frank, L. D. (2006). Neighborhood environment walkability scale: Validity and development of a short form. *Medicine & Science in Sports & Exercise, 38*(9), 1682-1691.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6*(1), 57-79.

Cheung, M. W. L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology, 37*(5), 522-541.

Cizek, G. J. (2008, April). Standard setting in the context of augmented achievement tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Occasional Paper of the Stanford Evaluation Consortium, Stanford University. (ERIC Document Reproduction Service No. ED 135 801).

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12*(1), 45-57.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly, 16*(1), 149-167.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*(10), 640-647.

Farmer, G. L. (2000). Use of multilevel covariance structure analysis to evaluate the multilevel nature of theoretical constructs. *Social Work Research, 24*(3), 180-191.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105-151). New York: Macmillan.

Goldstein, H. (1998). MLwiN release software. London: Institute for Education.

Goodwin, L. D. (1999). The role of factor analysis in the estimation of construct validity. *Measurement in Physical Education and Exercise Science,* 3(2), 85-100.

Hall, R., Hanges, P. J., & Dyer, N. (in press). Two-level structural equation modeling with latent variables: An illustration of confirmatory factor analysis with applications to organizational research. *Organizational Research Methods.*

Härnqvist, K., Gustafsson, J. E., Muthén, B. O., & Nelson, G. (1994). Hierarchical models of ability at individual and class levels. *Intelligence, 18,* 165-187.

Heck, R. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89-127). Mahwah, NJ: Lawrence Erlbaum.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques.* Mahwah, NJ: Lawrence Erlbaum.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hox, J. (2002). *Multilevel analysis.* Mahwah, NJ: Lawrence Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science, 39*(1), 1-22.

Joint Committee on Standards for Educational Evaluation (1981). *Standards for evaluations of educational programs, projects, and materials.* New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL for Windows.* Lincolnwood, IL: Scientific Software International, Inc.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: Praeger.

Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22*(3), 323-347.

Kaplan, D, & Kreisman, M. B. (2000). On the validation of indicators of mathematics education using TIMSS: An application of multilevel covariance structure modeling. *International Journal of Educational Policy, Research, and Practice, 1*(2), 217-242.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Knapp, T. R. (1977). The unit-of-analysis problem in applications of simple correlation analysis to educational research. *Journal of Educational Statistics, 2*(3), 171-186.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Kuhlemeier, H., van den Bergh, H., & Rijlaarsdam, G. (2002). The dimensionality of speaking and writing: A multilevel factor analysis of situational, task and school effects. *British Journal of Educational Psychology, 72,* 467-482.

Li, F., Duncan, T. E., Duncan, S. C., Harmer, P., & Acock, A. (1997). Latent variable modeling of multilevel intrinsic motivation data. *Measurement in Physical Education and Exercise Science, 1*(4), 223-244.

Linn, R. L. (2006). Following the *Standards:* Is it time for another revision? *Educational Measurement: Issues and Practice, 25*(3), 54-56.

Linn, R. L. (2008). *Validation of uses and interpretations of state assessments.* Washington, D.C.: Council of Chief State School Officers.

Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury.

Luppescu, S., Gladden, R. M., & Bryk, A. S. (2003, April). Reconsidering reliability in a multi-level context. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 681-699). Westport, CT: Praeger.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading, 9*(2), 85-116.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics, 8*(4), 271-288.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, B. O. (2002, February 20). Cluster size [Msg 2]. Message posted to www.statmodel.com/discussion/messages/12/164.html?1191440281

Muthén, B., & Asparouhov, T. (2009). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. To appear in J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis.* UK: Taylor and Francis.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267-316.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Muthen, L. K. (2008, September 26). Twolevel EFA example [Msg 2]. Message posted to http://www.statmodel.com/discussion/messages/12/3591.html?1226512296

Muthén, L. K., & Muthén, B. O. (2008). MPLUS version 5.2. Los Angeles, CA: Muthén & Muthén.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). A multilevel approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport & Exercise Psychology, 26,* 90-118.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). HLM 6: *Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(3), 199-213.

Reise, S. P., Meijer, R. R., Ainsworth, A. T., Morales, L. S., & Hays, R. D. (2006). Application of group-level item response models in the evaluation of consumer reports about health plan quality. *Multivariate Behavioral Research, 41*(1), 85-102.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126-136.

Research Triangle Institute. (1994). *Software for survey data analysis (SUDAAN) version 9.0.* Research Triangle Park, NC: Research Triangle Institute.

Rumberger, R. W., & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *Sage handbook of quantitative methodology for the social sciences* (pp. 235-258). Thousand Oaks, CA: Sage Publications.

SAS Institute, Inc. (2005). *The SAS system for Windows, version 9.* Cary, NC: SAS Institute, Inc.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analyses. A Festschrift for Heinz Neudecker* (pp. 233-247). London: Kluwer Academic Publishers.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 399-419). Thousand Oaks, CA: Sage.

SPSS, Inc. (2006). *SPSS for Windows,* Rel. 14.0.2. Chicago: SPSS, Inc.

Sexton, J. B., Helmreich, R. L., Neilands, T. B., Rowan, K., Vella, K., Boyden, J., Roberts, P. R., & Thomas, E. J. (2006). The Safety Attitudes Questionnaire: Psychometric properties, benchmarking data, and emerging research. *BMC Health Services Research, 6,* 1-12.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323-355.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement, 17*(4), 245-282.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237-259.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Stanat, P., & Lüdtke, O. (2008). Multilevel issues in international large-scale assessment studies on student performance. In F. J. R. van de Vijver & D. A. van Hemert (Eds.),

Multilevel analysis of individuals and cultures (pp. 315-344). New York: Lawrence Erlbaum Associates.

Steele, F., & Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of women's status. *Sociological Methods & Research, 35*(1), 137-153.

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Tate, R. L., & King, F. J. (1994). Factors which influence precision of school-level IRT ability estimates. *Journal of Educational Measurement, 31*(1), 1-15.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind.* Chicago: The University of Chicago Press.

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*(2), 272-296.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Van de Vijver, F. J. R., & Watkins, D. (2006). Assessing similarity of meaning at the individual and country level: An investigation of a measure of independent and interdependent self. *European Journal of Psychological Assessment, 22*(2), 69-77.

Van den Berg, S. M., Glas, C. A. W., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics, 37*(4), 604-616.

Van Horn, M. L. (2003). Assessing the unit of measurement for school climate through psychometric and outcome analyses of the school climate survey. *Educational and Psychological Measurement, 63*(6), 1002-1019.

Van Peet, A. A. J. (1992). *De potentieeltheorie van intelligentie.* [The potential theory of intelligence] Amsterdam: University of Amsterdam, Ph.D. Thesis.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Washington, DC: Council of Chief State School Officers.

Westat. (2000). *WesVar Complex Sample 4.0*. Rockville, MD: Westat.

What Works Clearinghouse (2006). Evidence standards for reviewing studies: Revised September 2006. Downloaded on December 17, 2006 from www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Yammarino, F. J., & Dansereau, F. (Eds.). (2004). *Multi-level issues in organizational behavior and processes.* San Diego, CA: Elsevier.

Zhang, N. J., & Wan, T. T. H. (2005). The measurement of nursing home quality: Multilevel confirmatory factor analysis of panel data. *Journal of Medical Systems, 29*(4), 401-411.

Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem scale. *Educational and Psychological Measurement, 65*(3), 465-481.

Zumbo, B. D. (2008, April). Testing and measurement from a multi-level view: Psychometrics and validation. Paper presented at a joint session of the annual meetings of the American Educational Research Association and National Council on Measurement in Education, New York, NY.

Zumbo, B. D., & Forer, B. (in press). Testing and measurement from a multilevel view: Psychometrics and validation. To be published in J. Bovaird, K. Geisinger, & C. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K-12 settings [festschrift to Barbara Plake]*. Washington, D.C.: American Psychological Association.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12*(2), 127-140.